DOCUMENT RESUME

ED 243 891 TM 820 280

AUTHOR	Schwartz, Judah L., Ed.; Garet, Michael S., Ed.
TITLE	Assessment in the Service of Instruction.
INSTITUTION	Massachusetts Inst. of Tech., Cambridge. Div. for Study and Research in Education.
SPONS AGENCY	Ford Foundation, New York, N.Y.; National Inst. of Education (ED), Washington, DC.
PUB DATE	Jan 81
GRANT	G-79-0045
NOTE	167p.
PUB TYPE	Collected Works - Conference Proceedings (021) -- Viewpoints (120)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Criterion Referenced Tests; Educational Assessment; Educational Diagnosis; Elementary Secondary Education; Instruction; *Instructional Improvement; *Instructional Materials; Learning Processes; Measurement Techniques; Standardized Tests; Testing; *Tests; *Test Use; Test Validity
IDENTIFIERS	*Alternatives to Standardized Testing

ABSTRACT

In an effort to examine issues raised by the effort to assess the performance of educational institutions, a project focusing on the social purposes and intellectual foundations of assessment practices in education was initiated. The primary goal of the project was to explore the possibility of developing new, more appropriate educational assessment strategies. As part of the project, several panels were convened, each focusing on a broad purpose of educational assessment. This document is the report of the first panel, which focused on the role of assessment in classroom instruction. It includes papers by Eva L. Baker, Eugenia Kemble, Philip Jackson, David Hawkins, J. Parker Damon, Asa G. Hilliard III, Howard E. Gruber, Robert T. Keegan, Judah L. Schwartz, Edwin F. Taylor, Nancy Willie, and Michael S. Garet. The panel concludes with four recommendations: (1) in developing new assessment materials, it is worth starting small; (2) the development of new assessment materials should be carried out by groups with a strong interest in the content areas being assessed; (3) schools interested in adopting new forms of assessment should begin by focusing on a small number of classrooms and subject areas; and (4) making new forms of assessment work in practice will depend on the sensitivity and ingenuity of teachers. (BW)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

G-79-0045

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED243891

ASSESSMENT IN THE SERVICE OF INSTRUCTION

edited by

JUDAH L. SCHWARTZ
Massachusetts Institute of Technology

and

MICHAEL S. GARET
Stanford University

The report of a study panel to
THE FORD FOUNDATION and
THE NATIONAL INSTITUTE OF EDUCATION
JANUARY 1981

TM 820 280

WORKING DRAFT
copyright 1981 - MIT

TABLE OF CONTENTS

PREFACE

INTRODUCTION............................Judah L. Schwartz & Michael S. Garet

Part I. Assessment and Instruction; What Is

TESTS & INSTRUCTION; AN HISTORICAL OVERVIEW....................Eva L. Baker

TEACHERS NEEDS FOR INFORMATION................................Eugenia Kemble

Part II. Assessment and Instruction; What Might Be

THE UNCERTAINTIES OF TEACHING................................Philip Jackson

THE UNPREDICTABLE NATURE OF LEARNING.........................David Hawkins

Part III. What To Do and What To Avoid Doing

INVESTIGATIVE TEACHING.....................................J. Parker Damon

CULTURAL VARIATION & LIVING ASSESSMENT.................Asa G. Hilliard III

FOSTERING INTELLECTUAL DEVELOPMENT......Howard E. Gruber & Robert T. Keegan

Part IV. Some Steps Now Underway

CATEGORICAL TEST VALIDATION;
THE TORQUE APPROACH.............J.L. Schwartz, E. F. Taylor & N. Willie

Part V.

CONCLUSION & RECOMMENDATIONS...........Judah L. Schwartz & Michael S. Garet

PREFACE

Programs designed to assess student achievement and the performance of educational institutions have become widespread in American education. But as assessment programs have multiplied, doubts about their effectiveness have grown. Serious questions have been raised about the methodology of standardized educational testing. Disagreement has mounted over the interpretation of test scores. And controversy has arisen over the ways in which tests are used in schools and universities.

School boards, departments of education, legislatures, and the courts have all been involved in the debate over testing policy and practice. Large numbers of school districts have implemented testing programs designed to diagnose student progress in the basic skills. Some states have enacted laws requiring that students pass "minimum competency tests" for promotion from grade to grade. Other states have enacted laws requiring that students achieve a minimum test score in order to receive a high school diploma.

As of this writing, the legislature in one state has enacted a "Truth in Testing" law requiring that all standardized tests used in the state for university admissions decisions be made public after they are administered, and similar laws are being considered by other states and the United States Congress. At the same time, a U.S. Court in one state has ordered a moritorium on the state's minimum competency testing program. And a U.S. Court in another state has ordered that IQ tests no longer be used to place children in classes for the educable mentally retarded.

Altogether, the effort to assess the performance of educational institutions has raised a number of difficult issues|

* What roles should assessment programs play in educational policy and practice?

* What kinds of assessment materials are appropriate for these roles?

* What should be expected of educational tests, when they are used?

* What should be taken into account, in distinguishing appropriate and inappropriate uses of educational assessment?

In an effort to examine these questions, the Division for Study and Reseach and Education at M.I.T., with the support of the Ford Foundation and the National Institute of Education, has initiated a project focusing on the social purposes and intellectual foundations of assessment practices in education. The primary goal of the project is to explore the possibility of developing new, more appropriate educational assessment strategies.

There is a growing critical literature on the role of educational testing in the schools. Much has been written on the defects in currently available standardized tests. But little has emerged on alternatives to present practice. Thus, the aim of our project is a synthetic one, to search for positive guidelines for new approaches to educational assessment.

One of the main assumptions underlying our work is that there are a number of distinct social purposes educational assessment is expected to serve, and each of these purposes may require somewhat different assessment methods, instruments, and practices. First, schools conduct assessments to obtain feedback on student progress, so teaching methods and materials can be adjusted appropriately. Second, educational institutions conduct assessments as the basis for reports to parents, school boards, and government agencies, as a way of promoting accountability. Finally, schools and universities conduct assessments to provide information on whether a student has mastered a body of knowledge or skills (for purposes of awarding a diploma or license), or to provide information on how well a student will do in the future (for purposes of selecting applicants for colleges and professional schools).

Generally, these distinct social purposes have all been addressed using the same sorts of assessment instruments. There is little a priori reason to believe, however, that instruments designed to serve one of these purposes are equally suited to serve the others. Indeed, it seems more likely that the opposite is true. Consequently, we have organized our project by examining each of several of the social purposes of assessment in turn and asking what types of instruments and practices might best serve each if the constraints of present practice, tradition and vested interest were absent.

As part of our project, we have convened several panels, each focusing on one of these broad purposes of educational assessment. This document is the report of the first panel, focusing on the role of assessment in classroom instruction.

In forming the panel on instruction, we brought together people with diverse perspectives on education, and asked them to think broadly about the role educational assessment might play in the teaching and learning process. The members of the panel were:

Eva Baker (Director, Center for the Study of Evaluation, University of California at Los Angeles)

J.Parker Damon (Principal, McCarthy-Towne School, Acton, Massachusetts)

Howard Gruber (Professor of Psychology and Director, Institute for Cognitive Studies, Rutgers University)

Walt Haney (Senior Research Associate, The Huron Institute)

David Hawkins (Professor of Philosophy, University of Colorado)

Asa Hilliard III(Dean, Schol of Education, San Francisco State University)

Philip Jackson (Professor of Education, University of Chicago)

Robert Keegan (Rutgers University)

Eugenia Kemble (Special Assistant to the President, American

Federation of Teachers)

Carmen Perez, New York State Department of Education

Edwin Taylor (Education Development Center)

Sheldon White (Professor of Psychology, Harvard University)

Nancy Willie (Education Development Center)

Jerrold Zacharias (Professor Emeritus, Massachusetts Institute of Technology).

The Panel on Assessment and Classroom Instruction met for the first time in March of 1979. Over the next year, members of the Panel prepared outlines, comments, and draft papers, which were circulated and discussed at a second Panel meeting held in February of 1980. The Panel completed its work in June of 1980.

This document while reflecting the panels views is woven together with only the lightest of threads. The individual authors are in no way to be held responsible for the coherence the editors have not been able to make sufficiently explicit.

We wish to thank Lewis Pike of the National Institute of Education and Marjorie Martus of the Ford Foundation for the encouragement they have offered us in this work.

It is the editors' pleasure to acknowledge the assistance and good humor of Ligia Domingo in the preparation of the manuscript.

Judah L. Schwartz
Michael S. Garet

Cambridge, Mass.
Stanford, Calif.
January 1981

CHAPTER 1

INTRODUCTION

Since the turn of the century, educators have hoped that the practice of classroom teaching might be reformed through the use of standardized educational tests. As the discipline of psychological measurement took form in the first few decades of the twentieth century, practitioners believed that achievement tests might have a significant and beneficial effect on teaching. Tests might help teachers make more objective judgments about student progress. They might offer diagnostic information on student learning problems. And they might assist teachers in devising and evaluating instructional strategies.

There is growing doubt, however, about whether conventional standardized tests have provided much support for the classroom teacher. Many observers of educational testing argue that the tests commonly in use fail to provide information useful in the practice of teaching. Indeed, some observers argue that of the primary purposes tests are expected to serve — instruction, accountability, selection, and licensure — tests serve instruction least well. For example, the report of a recent National Institute of Education "Conference on Research on Testing" concluded: " Instructional guidance is the educational activity which is least served — some published articles have said not at all served — by existing tests." (NIE 1979)

Serious questions about the instructional value of tests have been raised by several recent studies of the role of educational testing in the classroom. One study, conducted by the Center for the Study of Evaluation at UCLA, found that teachers rarely use standardized tests to guide instruction. Instead, they rely on tests primarily to confirm judgments about students made in other ways.(Yeh 1978) Another study, conducted by the University of Pittsburgh, found that while teachers make frequent instructional decisions about individual students, they seldom if ever use test results as a basis for these decisions.(Resnick, Salmon-Cox & Sproul 1980) And in a survey conducted by the American Federation of Teachers, a majority of the teachers surveyed reported that tests do not provide sufficient information on instructional materials and activities.(Kemble, this volume)

Given the questions raised concerning the instructional value of conventional testing, we have set out to reconsider the role of assessment in the classroom. In particular, we have focused on the following problem: How can assessment strategies be devised to provide information helpful in the teaching and learning process? What assessment strategies would support instruction in the classoom?

In addressing these issues, we have been led to a view of the role of assessment in the teaching and learning process which differs in significant ways from the conventional view of the instructional uses of testing. In the literature on testing, it is possible to identify two somewhat distinct ways of thinking about the relationship between assessment and instruction. One approach grows out of the tradition of standardized psychological testing, and the other grows out of a more recent concern with learning theory and instructional objectives. Some of the ideas we will propose can be clarified by contrasting them with these two conventional views.(*)

One view of the relationship between assessment and instruction is based on psychological testing. Standardized psychological testing, of course, has a long history and a tradition of practice. Generally speaking, standardized tests are supposed to detect differences between individuals, with respect to stable, underlying traits or characteristics — such as visual memory or aptitude. Undoubtedly the most well-known standardized test is the "intelligence" or IQ test, which was originally developed to predict how well a child might do in school.

For our purposes, the most important standardized tests are the general achievement batteries, diagnostic tests, and readiness tests. Standardized achievement batteries are widely used in the elementary grades, and they are designed to compare student performance in broad educational subject areas such as reading, arithmetic, spelling, and language usage. Diagnostic and readiness tests are supposed to provide somewhat more specific information on a student's strengths and weaknesses in an instructional area. A third grade diagnostic reading test, for example, might provide scores on auditory vocabulary, auditory discrimination, phonetic analysis, structural analysis, and comprehension.

Standardized tests are thought to be useful in instruction because of a belief in their ability to predict future performance. For example, reading readiness tests are often used at the end of kindergarten or the beginning of first grade to predict which children will have difficulty learning to read. And standardized achievement tests are used to group children for instruction, under the assumption that children with similar scores have similar instructional needs.

The intended role of standardized tests in instruction is somewhat similar to the role of diagnostic tests in clinical medicine. Both educational tests and medical tests are used because they are expected to be good predictors. Medical tests are used, of course, because they are helpful in predicting the presence or absence of disease. Similarly, educational tests are supposed to predict the presence or absence of learning problems.(1)

The "medical model" of educational testing suffers from one main defect. Unlike medical tests, current standardized tests provide little information useful in what might be called "differential diagnosis." In spite of the fact that some tests are labelled "diagnostic," they generally provide little specific guidance about a student's strengths and weaknesses. While standardized tests sometimes predict student performance, they rarely help explain why students perform as they do.

This defect in the "medical model" of educational testing may simply indicate that researchers have not yet been able to identify and measure the underlying traits that influence learning. Or, the defect may be more serious. Perhaps, as some members of our panel argue, the notion of measurement, borrowed from the physical sciences, is inappropriate when applied to human talents and abilities.(Schwartz, Taylor & Willie, this volume)

In the last twenty years, a second, somewhat distinct view of the role of assessment in instruction has emerged, drawing in part on experimental learning theory. In this view, tests should be designed, not to detect individual differences on underlying traits, but rather to assess student progress toward explicit instructional objectives. This emphasis on instructional objectives is shared by a number of recent educational innovations, including programmed instruction, criterion-referenced testing, domain-referenced testing, and mastery learning.

To develop an objective-based test in a particular subject area, it is necessary to divide the subject area into appropriate instructional units or domains. One common way of doing this is to postulate a sequence of instruction, leading from lower-level to higher level skills. Once a sequence of objectives is established, the role of assessment in instruction is straightforward. Students are tested at the beginning of each instructional unit, to determine the areas in which instruction is required, and at the end of each unit, to assess mastery.

Although the movement to develop objective-based tests is still young, several questions can be raised about the instructional value of the objective-based tests currently in use. First, the effort to divide subject areas into sequences of objectives and sub-objectives often results in systems that are extremely large. The Individualized Mathematics System for example, an objectives-based arithmetic curriculum for grades 1-6, involves 393 objectives, organized into 11 content areas and 9 levels of difficulty. Because of the size of systems like these, they are often difficult to integrate with other instructional materials and activities.

Perhaps more important, the division of subject areas into instructional domains often seems arbitrary, especially for objective-based tests that are not linked to particular curricula. In general, little empirical work with children has been done to determine whether the instructional domains that have been carved out have any instructional significance.

Finally, instructional objectives systems often emphasize rote skills at the expense of conceptual understanding. The division of subject areas into small units often produces an arid, if not atomistic or reductionist conception of knowledge.

In summary, then, there are two popular views of the role of assessment in instruction. One approach has involved attempting to integrate standardized tests and instruction, through a model somewhat similar to medical diagnosis. The other approach has involved attempting to integrate objective-based tests and instruction, by organizing instruction in small, discrete units, so that tests can be inserted along the way.

We believe that there is another way of thinking about the relationship between assessment and instruction, an approach that may be more helpful in developing useful assessment materials. Rather than beginning the discussion of assessment by asking how tests should be developed, we think it is more helpful to ask how teachers, in their ordinary day-to-day classroom experiences, figure out what their students know.

In the act of teaching, after all, teachers continuously ask questions and make judgments. Teachers continuously ask themelves whether this child understands a particular concept, whether that child should spend more time in reading rather than social studies, whether this lesson is "getting across," whether that lesson is moving too rapidly or too slowly. Teaching itself is a continous process of inquiry — in other words, of assessment.

We believe it is helpful to view formal assessment materials — tests-- as ways of expanding upon the inquiry process already inherent in teaching. Assessment materials, in this view are not something separate from instruction, something to be used before or after instruction, but are instead something which is a continuous part of the act of teaching itself.

From this perspective, assessment materials might be conceived as materials much like regular classroom exercises, tasks, and games — but designed to provide a bit more information about how a student is thinking and what a student understands. Assessment materials should provide teachers a way of looking carefully at a student's regular classroom work, to see why the work was done the way it was.

Assessment materials of this kind might help teachers and students in several ways. First, they might help a teacher find pattern and order in the strengths and weaknesses appearing in a child's work. For example, a teacher might notice that a particular child has difficulty forming plural nouns, and an assessment exercise focusing on plurals might call attention to some potential sources of the problem. Another child might perform erratically on arithmetic word problems, and an assessment game might help determine which sorts of word problems are causing difficulty and why.

Assessment materials of this kind might also serve another purpose. They might help teachers communicate with each other about individual children and their work. For example, such assessment materials might provide well-focussed, concrete examples of a child's work, so that teachers can discuss problems and suggest solutions. In the same way, assessment materials might help teachers communicate with parents about specific strengths and weaknesses.

Altogether, assessment strategies of the type we are proposing would have three main characteristics. First, they would help identify regularities underlying the strengths and errors in children's work. Second, they would respect diversity among children, and they would draw on children's life experiences in their own culture. Third, they would serve as the basis for dialogue — among teachers, students, and parents.

In the report that follows, we develop this alternative view of the role of assessment in instruction in some detail. The report contains five parts. In Part I, we discuss the problems teachers face in trying to use currently available testing materials. In Part II, we develop some of the main philosophical themes underlying our view of asssessment and instruction. In Part III, we draw on these themes to outline some of the characteristics we believe new assessment materials should possess, and in Part IV we describe a project whose aim was to develop assessment practices that embody some of the ideas discussed in Part IV. Finally, in Part V, we offer some recommendations for the development of new assessment materials, we consider what it might cost to move in the directions we describe, and we suggest some organizational and poltical strategies that might promote the practices we propose.

REFERENCES

Lauren Resnick, Leslie-Salmon Cox, and Lee Sproul, THE SOCIAL FUNCTIONS OF EDUCATIONAL TESTING, University of Pittsburg, 1980.

Ralph W. Tyler and Sheldon H. White, TESTING, TEACHING, AND LEARNING: Report of a Conference on Research on Testing, National Institute of Education, Washington D.C., 1979.

Jennie Yeh, TEST USE IN SCHOOLS, Washington, D.C.: U.S. Department of Health, Education, and Welfare and National Institute of Education, 1978. One major study of the role of tests in the classroom has obtained results quite different from those obtained by Yeh and the others discussed above. Michael D. Beck and Frank P. Stetz, of the Psychological Corporation, conducted a large sample survey of teachers, and a substantial majority of teachers reported using tests for instructional purposes. One reason for the discrepancy in the results of these studies may be that the studies discussed above asked somewhat more specific questions about the role of tests in instruction than did Beck and Stetz. See Michael D. Beck and Frank P Stetz, "Teacher Opinion of Standardized Test Use and Usefulness," Paper present to the American Educational Research Association, San Francisco, April, 1979.

NOTE

---

(1) For example, new born-infants are often given a test that measures the level of blood phenylalinine, because the level of blood phenylalinine is associated with PKU, an inherited metabolic disease. Infants whose blood levels are above normal are more likely to have PKU than children with low blood levels.

Like educational tests, medical tests are often far less than perfect predictors. Not all children with high levels of blood phenylalinine, for example, actually have PKU. Babies who are premature sometimes show high levels of blood phenylalinine.

For a useful account of the predictive model and the role of diagnostic tests in medicine, see Robert S. Galen and S. Raymond Gambino, BEYOND NORMALITY: THE PREDICTIVE VALUE AND EFFICIENCY OF MEDICAL DIAGNOSIS, New York: John Wiley and Sons, 1975.

PART I

ASSESSMENT AND INSTRUCTION: WHAT IS.

Educational achievement testing is a familiar feature of elementary and secondary school life. Achievement tests are widely administered, and their results are periodically reported to teachers, school departments, parents, government agencies, and even, from time to time, local newspapers. We begin our discussion of educational testing with a series of questions. What kinds of tests are generally used in the schools? What assumptions about teaching and learning do these tests reflect? What role do the tests play in classroom instruction? And how well do they serve the teaching and learning process?

Eva Baker, Profesor of Education and Director of the Center for the Study of Evaluation at the University of California at Los Angeles, provides an historical overview of educational testing in the United States. Baker begins her paper by discussing the origins of standardized psychlogical testing and educational achievement tests. As Baker points out, standardized achievement testing grew out of an effort to identify individual differences on underlying traits, largely for purposes of educational prediction. Baker concludes her discussion of standardized testing by raising some serious questions about their instructional value.

In the last twenty years, a second approach to educational testing has emerged, partly as a result of criticisms of the intructional value of conventional standardized testing. Baker outlines the development of this second tradition — sometimes called objective-based or criterion-referenced testing — and discusses some of the difficulties involved in specifying instructional objectives and using criterion-referenced tests in the classroom. One response to these difficulties has been a recent shift of attention from instructional objectives to learning domains. Baker raises some questions about this recent trend, and then discusses some dilemmas that must be considered in the development of new, more useful assessment materials.

In the following chapter, we turn from an analysis of the assumptions underlying conventional educational tests to an examination of the role these tests play in the teaching and learning process. Eugenia Kemble, Special Assistant to the President of the American Federation of Teachers, reports the results of several surveys of teachers opinion about testing and draws some conclusions about the kinds of information teachers would like tests to provide.

According to a survey conducted by the AFT, teachers believe current tests provide insufficient guidance for instruction. The survey indicates that teachers desire assessment materials that provide more information about individual students, and especially about student strengths and weaknesses.

Kemble then outlines some characteristics assessment materials should possess, if tests are to support teachers in the practice of teaching. In particular, Kemble argues that the traditional distinction between assessment and instruction should be reconsidered. Assessment materials not only generate information about students. They also influence what is taught and what is learned. Thus, it is essential that educational testing materials reflect the depth and diversity of the aims of education.

CHAPTER 2

TESTS AND INSTRUCTION: AN HISTORICAL OVERVIEW

Eva L. Baker

University of California
at
Los Angeles

Educational testing provides accountability; testing raises standards and facilitates learning; educational testing proscribes teaching. These paradoxical interpretations of testing occur partly because of the operating understanding we have for the enterprise. Where do these understandings come from? What is an achievement test? What could count as a test? Do we know whether the tests we have are the tests we might have? Should tests be resurfaced, remolded, or retained? How did we get what we have in educational testing?

In common experience, tests have come to mean "trials", as in the trials by fire suffered by mythic heroes. Tests are endured because of the rewards they promise upon success. In the sense that mettle is tested and ability found out, tests are thought to have revelatory power. They investigate personal limits and secrets; they display what people are inside. Tacit acceptance of this revelatory potential is, in part, what makes people anxious about tests.

Tests and trials are also terms common to the language of judicial procedure. Legal "tests" create additional nuance for our definition, because courts are convened to discover the truth. In law, truth is to be determined fairly, and due process requires that particular rules of demonstration and evidence be followed.

Tests are also employed in medicine, to verify or exclude alternative causes of particular symptoms. In the realm of science, tests are used to examine the tenability of hypotheses. And in engineering and applied science, tests may be used in reaching critical decisions, (for example, to determine whether a machine falls within a band of acceptable performance) or they may function more simply as observation points within a carefully specified set of conventional procedures.

The word "test" has been woven into our most casual conversation, partly no doubt as a result of our fascination with technology, and with research and development. Test pilots, men who braved the dangers of new supersonic aircraft two decades ago, have now been reified by inversion: instead of people, we have events; "pilot tests" stand for the tryout of something under development, a trial which occurs under conditions of at least minimum verisimilitude.

Revelation, psychology, law, medicine, science, and technology are high-status sources of the connotations associated with testing. Undoubtedly, all these uses and understandings of "test" somehow contribute to the range of interpretation evident in educational testing and, as certain, all occur against a background socioeconomic system based on competition. But as influential as these connotations may be, the specific applications of testing in education require further exploration.

Our discussion narrows, then, to the uses of achievement tests in education. The principal uses of tests since their inception have been for *placement* (to decide who belongs in a particular class or instructional program) for *credentialing*, or grading (to determine who did how well, or who did well enough) and for *program evaluation* (to find out what changes are needed in educational sequences). While there are numerous other uses of tests, let us confine discussion to the three identified for achievement testing.

THE RISE OF STANDARDIZED TESTING

The relative emphasis given to placement, credentialling and evaluation has varied over the history of educational testing. In the early days of testing, placement or selection was paramount. One of the first standardized educational tests developed in the United States, for example, was used to select men to be officers in U.S. military service. Tests were developed so that they would detect individual differences among potential officers. The test development paradigm was parallel to that employed by Binet in his well known explorations of human intelligence.

The prevalent statistical models of the time reinforced the differentiation function of tests and provided comparative information about individual performance. Of high interest was whether an individual placed in the top, middle, or bottom of a distribution of scores. Because reliability, or the consistency of a person's rank in a distribution was needed, great emphasis was placed on the stability of ranking; a person who was best on one day should, when readministered the test, be best again, or close to it. Concommitant with this notion of test stability was the interpretation of human test performance as a measure of a *stable* characteristic or general *trait* possessed by the learner.

The importance of prediction cannot be overstated in this model. Philosophically, the model suggests that schooling operates to sort into groups people of various stable and predictable characteristics thought to profit *differentially* from alternative instructional regimens. One extension of this view can be discerned in recent research by those who want to match a student's instructional treatment with the student's cognitive style. This line of research is called, alternatively, trait-treatment interaction or aptitude-treatment interaction, and the "trait" or "aptitude" is usually measured by achievement tests. (Immediately, one should pereceive a basic conflict in this view of the function of testing: "achievement" is seen both as a predictable, stable trait and as something amendable to change, perhaps through schooling.)

Tests of achievement, developed originally in order to differentiate among individuals (including ubiquitous college entrance examinations), gained legitimacy from a number of sources. First, the tests promised to add an important refinement to selection processes, a refinement in the name of the democratic principle of fairness. It became socially less acceptable, although perhaps not less frequent, to select those for the educational elite exclusively from among the ranks of the wealthy. Tests seemed a fair way to broaden the information used in selecting students for higher education. Second, it was important to recognize that selection into programs such as university education or officer's candidate school was regarded for a long time as a special or uncommon reward and opportunity, not within the aspirations of most of the population. People voluntarily "sat for" college entrance examination; they were not required to do so. Thus, the tests were accepted as a legitimate tool to identify the deserving few. In general, those selected for college were rewarded (and were able to afford the option); those not accepted were not stigmatized or regarded as failures.

Concommitant with these social interpretations of testing was the continued development of procedural and statistical methods designed to support the sorting and placement purpose of tests. Before World War II, additional uses of achievement testing had not gained much importance. Except in particular professional or technical fields, and in the New York State Regents examinations, certification (the releasing of students from programs of instruction, or the passing of students from one program to another) was the private responsibility of academic personnel in schools. The teachers' right to assign grades was understood and generally unchallenged. Tests used to evaluate teaching and instruction appeared only sporadically with no concrete impact.

Since World War II, the role of standardized testing in education has expanded markedly. What forces account for this expansion? First, one may point to the democratization of schooling and the delivery of universal education. More and more students attended high school. Graduation become expected rather than exceptional. And through the effects of legislation designed to reward those with military serevice, college education became an economic possibility for a more diverse set of students. The student loan programs and the rapid growth and variation of the higher education system changed normative values, and students in increasing numbers planned to go and actually went to college.

While the expansion of schooling helped promote growth in the use of tests, undoubtedly, the single most identifiable influence in the post-war field of educational achievement testing was the federal government, in both its direct and indirect effects. The federal establishment supported educational research in the sixties on a scale unlike that experienced previously. Psychologists and educators, in their thrilled (or, at least cheerful) exploration of instructional and curricular variables almost exclusively depended upon the growing array of commercially available achievement tests. Education schools, began to shift from predominately teacher training "craft" centers to bastions of educational research. If much of the education research produced in the sixties was not heart stopping, nonetheless, the availability of federal research support grew (such as that offered by the Cooperative Research Act of 1963).

In parallel, the expansion of higher education, with the strict imposition of "publish-or-perish" criteria, fostered dependence upon the production of science-like educational research. Waiting to play their part were standardized achievement tests. The direct and final push establishing the legitimacy of achievement tests came from the great investment in federally inspired social and educational programs in the 1960's. In Title One, a program to improve the learning of disadvantaged students, in Head Start, a similar program from a different bureaucracy, the government required evaluations of student learning as a measure of program effect. Spawning a sub-speciality of still growing proportions, and aped as they were by both state programs and state evaluation requirements, federally mandated program evaluations fixed achievement tests as the criterion of choice for educational evaluation. If the most competent educational researchers, by and large, accepted such tests as adequate criteria to judge their theories with hardly a blink, one could similarly assume the suitability of these tests to determine program quality. Coleman used such tests to assess the policy implications of segregation, and Jencks continued that tradition.

Although there were early and vocal dissenters, the testing industry grew, encouraged by favorable governmental regulation, and supported, by and large by "experts" in universities. Percentile ranks, stanine scores, and reports of individual and school performance became commonplace; newspapers reported reading achievement scores, and school boards came and went on the strength of such scores. Poor "test" performance became a scandal, and test scores served as a marker for "educational quality". Repeatedly, conventional wisdom about the effects of one or another clearly different programs was swamped by test results tha showed no difference in achievement among differing concentrations of resources. It is only lately that the validity of test scores as measures of educational effect have been challenged.

So, advances in educational psychology, sunny optimism in the federal support of school interventions and research, the rise of higher education supported again by the government, and the blessing of achievement tests as satisfactory devices to measure educational growth conspired to create the climate of assessment we are faced with today. Particularly:

1. Tests of achievement evolved because of needs to choose the best students for special opportunities like college.

2. Tests were designed to measure stable student characteristics
.

3. The broadening of student populations and the value placed upon higher education contributed to the acceptance of standardized achievement testing.

4. The collective desire of educational research to approximate science, combined with the expansion of higher education and salient tenure decisions, inexorably increased experts dependence upon achievement tests for research studies.

5. Achievement tests were also used to certify students and to assess program efforts, although these uses came somewhat later.

6. Federal and state educational programs, requiring evaluation of innovations fed the growing testing industry.

7. Test scores, legitimated by these, and other well publicized and influential events such as the Coleman studies, created a climate where testing is seen as an essential component in educational programs and the single best indicator of educational quality.

STANDARDIZED TESTS AND INSTRUCTION

The tests used as "achievement" measures in the schools have mainly been commercially available, standardized tests, usually dealing with a subject matter or skill area, such as mathematics or reading comprehension. The fact that most achievement tests are both commercial and standardized, has had a large influence on the role of testing in instruction.

The commercial character of such tests has led them to be considered partly in terms of their marketability. Tests which are marketable are those with the broadest appeal, that is, those least tied to local or idiosyncratic needs. Thoughtful scholars in this field have pointed out that the requirement to be "general" and appropriate for broad use conflicts with the test's function to detect particular effects associated with identified programs (in program evaluation contexts.) Sampling a broad field, such as reading, with a test not sensitive to particular pedagogical methods can (and does) produce results that inaccurately portray achievement of students.

An even more insidious problem has been identified by Porter and his colleagues in Michigan in their analysis of mathematics and reading standardized tests. The technical manuals accompanying such tests describe the general topics to be tested, but review of the actual number of items used to assess different skills varies greatly and could selectively penalize classrooms of students whose instruction has not matched the same set of content emphases. Remedying the problem, that is, trying to select a test which better fits particular instruction is a course of action hampered by concerns for test security, a topic to be treated more extensively later.

The term "standardized" brings with it additional difficulties in test application. "Standardized" refers to at least two different, but interacting features of testing. One interpretation of "standardized" relates to the conditions of administration of the test. Wherever the test is given, a particular set of uniform directions is used; a specified amount of time is allocated; certain pencils may be required; common answer sheets can be provided; student questions about the test may (or may not) be answered; instructions to guess may or may not be given. The standardization, or exchangeability, of conditions of administration undoubtedly contributes to the test's special and ceremonial qualities. Such tests could not be given comfortably within the regular classroom daily life. Special rules are used and these rules contribute to the distinctiveness of the testing occasion compared with other classroom events. Atypical conditions likely affect anxiety most obviously for students, but with growing concern by teachers. The standardization of conditions, thus, underscores the foreigness of the test and sets it apart from "normal" instructional activities.

A second interpretation of the term "standardized" relates to the standardized way in which test results are to be interpreted. In common practice, the standards used to interpret test performance involve transforming raw scores (the number of right answers on a test) into formats that allow cross-student or cross-school comparisons. Test scores are most often scaled to produce a normal distribution, or what is sometimes called the bell-shaped curve. The reports of students' or schools' achievement levels are then converted to relative scores: A student might be in the first quartile (the bottom 25% of a distribution) or the 85th percentile, (with 85 percent of a comparably tested group scoring less well), or the fourth stanine, (placing the score somewhere in the middle of a set of scores.) Information is focused on a student's rank in a distribution rather than the student's level of skill or understanding.

This transformation of scores into a standardized framework for interpretation raises two related issues. First, how appropriate are the groups used to "norm" the tests? Norming groups may be inappropriate because of the socioeconomic characteristics of the students chosen, or because of the time elapsed since the norming process took place. For instance, should test scores from Minnesota in 1974 be compared with those in inner city Los Angeles in 1980?

Second, how does the effort to insure a normal distribution of test scores influence the relationship between testing and the curriculum? In order for this "normal" feat to occur, each students should have about 50-50 chance of getting each item right. But clearly, those items on which instruction has focused would have success levels much higher than 50%. Paradoxically, such items would be excluded from achievement tests for being too easy. And the paradox extends. Because tests are believed to measure stable traits, designing instruction to improve test scores (even if such could be done) is regarded as unethical if not perverted.

Thus, commercially available, standardized, norm-referenced tests create the potential for a logically strange set of conditions, if not simultaneously, at least in sequence:

> Tests that are commercial have to be general rather than specific, and yet they are expected to serve potentially idiosyncratic local program requirements;
>
> The administration of tests that are standardized requires some degree of foreign and special procedures, procedures which withdraw such tests from the day to day regularity of classroom instruction.
>
> Tests that are standardized using some form of the normal distribution provide relative information (who's better than whom) and they are most efficient when a student's chance of success on each item is about 50-50.

The general relationship to instructional programs of commercial, standardized tests is therefore weakened by: 1) general rather than specific content relationships; 2) the loss of information by providing relative data; and 3) the need to discard potentially instructionally relevant items to approximate the normal curve. Overall, then, the facts of development, administration and scoring of norm-referenced tests tend to weaken their utility for instructional planning, and therefore, result in reduced appropriateness of such measures as indicators of the effects of instructional programs. The process is unfailingly interactive. Notwithstanding, achievement tests continue to have wide influence in public education.

THE EMERGENCE OF OBJECTIVES-BASED TESTS

The linkage among achievement tests, academic psychology, and educational statistics, represents by no means the only approach that has been taken to the assessment of achievement. A contrasting perspective argues that tests should not be revered principally as instruments for measuring the "true capacity" of individuals, but should instead be seen as instrumental in the teaching and learning process itself.From this perspective constructs binding tests to curricular or instructional requirements have greater utility in achievement testing than do constructs of human capacity and individual differences. This view, dates from early in the 19th century, when Rice tested spelling performance,and it extends to the domain-referenced testing advocates of today.

Attention to the control of instructional events has led to a renewed interest in demonstrating the short-term effectiveness of instruction. Clear examples of this effort can be seen in the programmed instruction movement of the late 1950s and early 1960s. Programmed instruction was based on the idea that learning could proceed incrementally and, to some extent, diagnostically. "Programs", that is "reproducible sequences of instruction" were designed in order to control performance at every step of the way. Learner's responses were carefully monitored. Interrogations of "why" students may have responded one way or another way were the subject of research studies and recommendations.

The principles upon which these programs were based share some of the ideas promoted and debated by current test designers. For example, an early schism on response mode erupted between two major divisions of the field. The debate was whether it was best to ask respondents to construct or to select answers. One side favored gradually increasing the item difficulty of a learner-produced response; order of instruction was fixed, but students' time-to-completion, or rate, could vary with individual differences. The other side focused on multiple choice responses that built in attractive error options and appropriate remedial instruction contingent upon selection of different wrong answers. Thus, different students might experience very different presentations and orders determined by their error patterns.

Segments of instructional programs, called "frames, were designed to correspond to items on achievement tests. In a frame, the learner received a stimulus that presented the minimum number of features thought necessary to elicit a correct response. By careful trials, these frames grew gradually more difficult, so that at the end of the program the student was successful at comparatively difficult tasks.

The experimental psychologists who were the designers of programmed instruction were interested in making these sequences both effective and efficient. They contrasted prompted frames, where the learner received "help", with unprompted or criterion frames, where the learner performed the task unaided. Prompts, might be formal, e.g. line length or thematic and substantive, capitalizing on preexisting information possessed by the learner. With support from research, many psychologists believed that prompts should be "faded" or withdrawn, so that students became competent as quickly as possible. The term "lean program" was coined, suggesting that anything not demonstrably instrumental to performance should not be included in the sequence. While this emphasis on efficiency had a "Gee-whiz, look no hands!", appearance, the effect was to try to express, as concretely as possible, the particular set of performance tasks to indicate that "mastery" had been acquired.

On the issue of how one determined standards of either minimal or expert performance, most program designers chose pleasantly redundant numbers: 90-90, meaning ninety perecent of the students were to obtain ninety percent of the items correct, 80-80, and so on. Military training directors set high standards (that is, 90-90). This formal proclamation of standards (the program would need to be revised until the criterion level was set) sometimes resulted in the selection of easier tasks and the development of simpler items, so that 90-90, or whatever, was more practical to achieve.

Many well respected, present day scholars in cognitive science expended a good deal of their early scholarship detecting conditions under which programs were more successful and efficient. The effects of the spate of experimental work by these scholars and others was to legitimate a new form of achievement test, called in programmed instruction parlance, a criterion test, whose practical meaning was that the frames were unprompted. When Glaser first introduced the term criterion-referenced tests, his use of the term "criterion" was interpreted into two different ways: 1) "criterion" performance was the final or "terminal" set of tasks; 2) "criterion" referred not to the set of tasks or skills to be performed, but instead to the level of performance to be exhibited. Glaser's article stimulated a great deal of work attempting to extend his definitions. Yet the confusion between criterion set and criterion level remains, and it is reflected in many of the criterion- referenced tests developed.

What must be seen as most significant about the emergence of this testing framework as an alternative to standardized, norm-referenced achievement tests is the context from which such ideas emanated. Criterion testing, however flawed and imprecise at the outset, grew as a natural extension of instruction. These were tests to assess instruction, tests of specific and generally replicable teaching. The early definitions of programmed instruction emphasized its "reproducible" quality, and the term first stood for things that could be "dittoed" and later, Xeroxed, like paper and pencil, programmed booklets. The definition later expanded to include "a set of events ... essentially reproducible." Although the application of "criterion-test" to classroom instructional settings progressed rather slowly, and often with awkward control mechanisms (such as "scripts" for teacher-student interaction), in time, the idea that the teacher was principally responsible for instructional consequences was here-and-there acknowledged.

The translation process from "programmed" to teacher-led instruction stimulated a number of pertinent developments in the field of criterion-referenced testing. For one, the tendency to specify criterion tasks exhaustively, in the form of highly specific behavioral objectives, caught hold in public education for a brief time, and was even made statutory in some places. Some delight was found by those who enjoyed concrete experiences. For example, hundreds of reading objectives were identified for a single semester's instruction, and Michigan State University initiated a teacher training program with literally hundreds of objectives and tests specified.

A review of many health education programs demonstrates that behavioral specification still has a home. But two factors have diminished the zeal for behavioral objectives in all but the most fortified behaviorist encampments.: 1) the information load large systems of objectives place upon teachers; and 2) the nagging concern that objectives such as "The learner will be able to print a "j" have a relatively small set of items appropriate for measurement.

A concern for legitimating "criterion-referenced" measures stimulated individuals to attempt to apply some of the statistical approaches used in standardized testing to analyze test items. Such experiments in translating parametric statistical procedures for use with criterion-referenced tests point up some rather distinctive differences between the two test types. First of all, criterion-referenced tests are developed to be sensitive to instruction. Thus, following teaching, students' performance tends to clump near the high end of the scale, and before instruction, students' performance typically is arrayed at the bottom. The comparative lack of variation in scores first before and then after instruction has suggested that radically alternative statistical procedures are necessary.

Simultaneously, the criterion-referenced test advocates, largely from the instruction rather than psychometric side, continued to develop measures. The problem test designers had in common was how to describe the "criterion" tasks the tests were supposed to measure. The use of ordinary-language such as "to understand" or to "know" in describing tasks was ridiculed by many. Bloom's article on mastery learning both highlighted programmed instruction's expectations (i.e., instructor responsibility for learning) and accelerated this view of testing for teacher led, non-programmed contexts. Bloom had demonstrated that common cognitive processes could in fact be illustrated in many ways. Knowledge could be assessed by myriad test item formats; similarly so could "higher" processes on this posited taxonomy, including analysis and synthesis.

The behavioral psychology fervor of the instructional development groups did not countenance such "vagueness". Gagne, working principally to determine the structural relationships among learning components, suggested a framework that was more acceptable to those wishing a more concrete approach to task specification. He analyzed a set of five types of learning, and later proposed that any well stated goal should include a statement about the learners' cognitive process as well as a common format in which the performance was to be exhibited. The complementary work of Bloom and Gagne has encouraged the aggregation of cognitive tasks under common "levels" of learning, as one way of dealing with the glut of tasks, objectives and test items. Corresponding work in the specification of concept learning has also continued.

As methods of aggregation for objectives were developed using complexity of cognitive process as a heuristic guide, so other methods were explored to synthesize disparate objectives. In the early fifties Ralph Tyler identified "objectives" as consisting of two parts: behavior and content. The focus on specific behavior was first atomized by the compulsive specifiers and later understood and consolidated by the followers of Gagne and Bloom. But the content specification had yet to be systematically addressed within a measurement context. Although the specification of content-behavior matrices had been used by developers of standardized, commercial tests, in practice these specifications principally guided the initial versions of the test, and were less influential following empirical trial. For instance, if an item was found to be "too easy", the item might be revised to include more attractive "wrong" answers or to obscure the distinctiveness of the right answer. The bases for these decisions derived from the data rather than from any prescriptive notions about learning. Such revisions have often proved difficult to describe. In practice, the updating of test specifications has often seemed inconsequential once empirical data were collected. Test specifications, then, have often been used as a rough planning aid for item writers, rather than as rigorous guidelines. Indeed, item specifications have sometimes been written after the items themselves have already been prepared.

FROM OBJECTIVES TO DOMAINS

From the psychological learning perspective, the detailed specification of learning tasks collided with one of research's most cherished notions, i.e. the idea of generalization. In learning terms, the notion of "transfer" describes the "spillover" effects of instructional treatments, effects usually thought to be desirable. These side effects, for instance have often been included as dependent measures for research studies where performance on practiced and non-practiced tasks was contrasted. To take account of the idea of transfer in deliberate instructional planning, one should describe the categories of learning outcomes desired, and teach to these broader categories.

Foreshadowed by the work of Osbourne, Wells Hively set out to find a way to explicate content categories. Hively used set theory as a point of departure. He provided a model by which instruction could be matched to an identified set, or "universe", of content to be sampled by test items.

These "domain referenced achievement tests" solved a number of problems simultaneously. First, the formulation led the way to the specification of a limited and manageable number of tasks, in contrast to the proliferation of specific behavioral objectives. In this model, transfer tasks were incorporated as set members, to be intentionally addressed by both instruction and measurement, rather than left outside to enter as good luck might provide. These procedures also explicitly integrated content domains and behavioral requirements.

Hively's suggestion was very simple. He proposed that an "item form" or "shell" be created that encapsulated the behavioral requirements of tasks, explaining the kinds of stimuli to be presented, the conditions of exposure, and the manner of response desired. In addition, he demonstrated the specification of classes of content, to which instruction pertained, content which could be considered fair game for assessment and to which the learner's skills and understanding presumably generalized. Hively argued that the specification of domains should involved not only a description of content, but also a description of the contrasts or discriminations required to demonstrate that the learner understands the critical features of the tasks. These content limits should identify rules or guidelines for selection of content, e.g., "all pairs of two digit numbers," by enumeration, "all poems by Keats and Hopkins;" or comprehensive example, "all words found on page one of the Los Angeles Times".

The exercise of trying to formulate rules for the identification of content limits or boundaries has been somewhat frustrating. It has become clear, for instance, that analysis of structural relationships within certain disciplines has not been sufficient to permit the abstraction of sensible rules to define content universes. It has also proved to be comparatively easy to apply this process superficially so that content domans "look" as if they have been created.

The attempt to produce sets of representative items has confirmed the difficulties involved in carrying out Hively's program. The Hively approach has gone through a number of permutations, and simplified versions for teacher consumption have been developed, as well as more complex formulations to guide professional test designers. Recent work has focussed on a variety of issues, including questions about how big a domain should be, how many items should be sampled, what are the most important characteristics of items, and whether items from a common domain should be equally difficult.

Several models of the relationship between test domains and insruction have emerged. The Hively model, for instance, first asked teachers to generate exemplary instructional plans. By analyzing lesson features, Hively and his staff on the Minnemast project abstracted the behavioral and content features for assessment. Mutual adjustment between measurement and instructional people occurred, but the source of the domain was predominately teachers and instruction. Analogous processes were used by the Learning Mastery System developers. They painstakingly analysed text materials and then developed specifications and tests which sampled the content areas incuded in these texts.

A second model has focused on the specification of the domain prior to the development of instructional materials. In a curriculum development project in primary reading, for example, project staff developed domain specifications and then proceeded to develop instruction that representatively addressed the domain identified. A similar model, used in the Detroit Public Schools, specified the domains and then acquired or developed a wide range of instructional materials that might be used to address the domains. (These approaches are represented in figure 1, below.)

MODELS FOR RELATING TEST DOMAINS TO INSTRUCTION

	Closed systems (curriclar packages)	Open systems (ecletic instruction)
Domain-led Instruction		
Instruction-led Domain		

THE CURRENT DILEMMA

Domain specification, of course, is vulnerable to a number of charges. The first, and most critical, is the source of content and behavior specifications. Ironically, these questions are of most concern when the domain designers are very explicit about their selection rules. It is far easier to accept a statement such as "Words used in fourth grade reading texts" as a content description than abstractions such as "words of Latin derivation, not more than three syllables in length." The apparent legitimacy of using text books to define domains is, of course, partly attributable to our respect for real artifacts. The fact that fourth grade texts got published "must" imply that the words within them were reviewed carefully and found to be satisfactory. It is much easier to question why Latin as opposed to Greek derivations should be included in a domain, and why three syllables should be selected as the cut off point. Clearly, however, attributing legitimacy to the content in texts involves a degree of wishful thinking, and the source of the content limits in domain statements is in need of review.

Popham has sought to defend arbitrary specification of domains (why three examples in a paragraph and not four?) by invoking the power of collective human judgment. He urges people simply to judge and decide. He is quite possibly correct about the common state of the art, or even what may be possible in the near future. But an analysis of legitimate sources of test domains seems warranted, particularly as the power of tests to guide and perhaps direct curricular and instructional efforts seems to be increasing.

Ideally, a well developed set of specifications would derive from mature knowledge in diverse fields. For example, in selecting distractors for perceptual tasks, research on cue salience, color, size, position, would clearly influence the critical discriminations to be assessed, as well as the range of examples over which the performace should generalize. Thus, the knowledge developed from cognitive psychology would be one important source of information. Similarly, analyses ought to assure that the topics and illustrations identified in the domain are those upon which there is reasonable consensus. (What are the five characteristics of declining world powers? and who says so?) Third, domain specifications should attend to pedagogical knowledge. Content features are delimited by practical constraints in teaching as well as by our record in successfully developing certain classes of skills. Exhaustive analyses that produce domains no one can teach have only marginal interest.

Similarly, developmental psychology can provide cues about the capacities of children to assimilate various sorts of skills, as a function of maturation. One should be quick to recognize, however, that such perceptions may vary with the theory under which developmental patterns are hypothesized.

Another much overlooked area is the language of the test items, and its semantic and syntactic complexity. The issue of language needs to be addressed much more specifically than by simply reporting readability levels (however useful these formulas may be for longer discourse than typical test stimuli). Particularly as the concern for equity and cultural diversity in testing garners more attention, these linguistic features are likely to be increasingly influential, and analysis of linguistic issues should be incorporated systematically in developing domains.

Assumptions about the ways children confront test materials should also be examined. Analyses of test "frames" from cognitive and linguistic perspectives might provide additional ways to design test and instructional items that respect diversity among students.

Tests are not going to go away soon, even those that we might personally regard as irredeemable. We should therefore look for tests to become more useful and more fair. Their uses in instructional planning may ultimately call for the true merging of instruction and testing. Tests no doubt should become increasingly available or public, because of the constitutional guarantees inherent in society's requirements to sort and choose people, for schools, for certification, for retirement or dismissal, and for additional educational experience. Tests need to be made more cheaply and in more variety. Solutions are needed to a host of technical problems, such as how standards are set, how long tests should be, and how much error we can tolerate in our decisions. Finally, we need to find better ways to assess several critically important competencies currently given too little attention -- such as speaking, writing, and thinking.

TEACHERS NEEDS FOR INFORMATION

by

EUGINIA KEMBLE
Special Assistant to The President
American Federation of Teachers

Attendance at highly publicized conferences on educational testing, or perusal of the literature of many national education organizations, would lead the average observer to conclude that teachers, and the education community in general, are highly suspicious of the modern testing industry and even hostile to the administration of standardized tests. It is unfortunate that this public relations-style warfare has eclipsed any examination of the real needs and concerns of those who use tests. Even more distressing is this debate's effect in creating an atmosphere which questions the value of comparative standards, and which undermines whatever resources and inclination we might have to improve upon the assessment methods we now have.

While many groups and organizations are expending vast amounts of energy attacking the testing industry, most teachers in classrooms have information needs related to test results that are now, for the most part, going unmet. Their basic commitment to testing does not mean they are uncritical. But they are not demanding that all student evaluations be subjectively administered by teachers who have developed the tests themselves. They are not calling for a moratorium on the use of all standardized tests. And they are not hostile to the use of minimum competency tests.

From what we in the American Federation of Teachers, AFL-CIO have been able to find out through our own survey work and by examining the work of others, teachers feel a need to know more about their students. They want to understand students' individual educational needs better, and they want to be able to compare their progress with other students. They want to use this information to improve upon what they do in the classroom. And, they believe that standardized tests are useful in providing them with some of this information. They want more, not less, information from tests, and they would like more in the way of inservice training to help them in interpreting and using tests results. In short, teachers are looking for more action-oriented information about their students because they want to be more effective teachers.

In discussing how to get from where we are to where we should be, it makes sense to begin with a discussion of how teachers themselves view the current situation. We can then begin to analyze the gap between what is wanted and the adequacy of what is available. Finally, it should be possible to speculate on what needs to be done to improve things.

HOW TEACHERS THINK ABOUT AND USE STANDARDIZED TESTS

This discussion will rely on two studies of teacher views of the uses of standardized tests. The first is an as yet unpublished survey done for the

American Federation of Teachers by the Center for Study of Evaluation at the University of California in Los Angeles. The second is a study entitled "Test Use in Schools" by Jennie P. Yeh, also of the Center for the Study of Evaluation. The results of both are supportive of one another even though sampling and methodology were quite different.

The AFT Needs Assessment Survey was sent to a stratified, random sample of 800 AFT members. The return rate was 19%, or a total of 153 questionnaires. The return was divided roughly evenly between elementary and secondary school teachers. Most were from urban and suburban communities. While the return rate is low, the fact that the conclusions of the AFT survey are reinforced by the results of the work by Beck and Stetz would seem to indicate tha the AFT work has validity.

It must be made clear that the AFT survey measured teachers perceptions of their own capabilities and needs. When respondents claim to know which types of decisions are best fed by information from aptitude tests and which from achievement tests, there is no way of knowing exactly what their assumptions and knowledge really is. It is perceptions, not absolutes that we are looking at. It is also safe to assume that those who answered the survey questionnaire were probably those who felt most self-assured about their own abilities and opinions.

The most interesting aspect of the AFT survey's conclusions for purposes of this discussion has to do with teachers use of test results. Student placement or grouping and diagnosis of individual student needs were the uses ranked highest among respondents. Secondary school teachers tended to rely less on this information for these decisions than did primary school teachers. Those teachers with no formal training in tests and measurement gave more weight to test results in making these decisions than those with greater expertise.

Ironically, despite the usage of these tests for these purposes, teachers also complain that standardized tests do not provide enough information in areas that would seem to be directly related. For example, 64% of the respondents said that "results do not provide prescriptive information, e.g., guidance as to what materials, instructional activities are needed." Over half (54%) found that "results do not provide an adequate profile of student strengths and weaknesses." These two were among the top problems related to test usefulness for teachers.

These results would seem to indicate that teachers recognize the shortcoming of standardized tests when it comes to making decisions they simply must make. Lacking other information, they use these tests anyway. This dilemma is further indicated by the fact that 64% criticize standardized tests for being inadequate when it comes to instructional planning. And yet, 31% percent of the respondents reported high expertise in doing precisely this with test results, a larger percentage than for any other test-related activity.

The AFT survey also asked teachers what a perfect test would provide. Their answers are strongly supportive of what the other survey results relate. TEACHERS NEED MORE INFORMATION THAT WILL HELP THEM KNOW INDIVIDUAL PUPILS BETTER.

One caution must be introduced at this point. The fact that teachers recognize the shortcomings of standardized tests for purposes of instructional planning, and yet at the same time use these tests for precisely that purpose, need not cause us to conclude that standardized test use should stop. It may be that we should develop new tests to satisfy teacher needs more precisely. But, such advocacy says nothing negative about the relevancy of standardized tests for OTHER purposes relating to group and student comparisons. (It would be a mistake to translate a teacher demand for more test information specifically geared to decisions about individual students into a call for throwing out standardized tests. The classroom uses of assessment and the broad policy uses of assessment are different and should not be confused. The fact that teachers need more test information for classroom use does not make the information necessary for policy uses — information more likely to be gleaned from standardized tests — irrelevant or invalid.)

The study by Jennie Yeh covered a sample of teachers in nineteen California elementary schools. 260 teachers returned useable questionnaires, a return rate of about 60%. One of the main findings of the study is that, while teachers often use standardized test results for placement and grouping of students at the beginning of the school year, they seldom use test scores to guide instruction throughout the year. Instead, they rely on other sources of information. (See Table 6, below.) According to Yeh,

> Teachers reported that of several possible sources of information, they most frequently used information from interactions with or observations of students, informal assessment techniques (e.g., oral quizzes, reading aloud) or results from teacher-developed tests to assess their students throughout the year. The least frequently used sources of information were the results from standardized and instructional program or curriculum embedded tests, while moderate use was made of information about students' place in a book and work assignments.(1)

About 53% of the teachers who responded to the questionnaire reported that they developed their own in-class tests. According to Yeh, teachers who developed their own tests reported that the most important reason for doing so was that their own tests "more accurately assess the effects of their instruction. In other words, their own tests were seen as content valid."(1) Teachers also reported that the format and wording of their own tests seemed more suitable for students. (See Table 7.)

There are some rather simple conclusions that derive from the data presented here. First of all, it is clear that teachers want and need assessment information. It seems, however, that most of what they are getting from standardized tests is not as useful to their decision-making needs as it could be. They realize this, but they often use the information anyway, an outcome that could be counterproductive. This would seem to indicate that more needs to be done to help teachers differentiate between test types and their valid uses. It also means, and this is the most important conclusion, that MORE TESTS NEED TO BE DEVELOPED TO SPECIFICALLY HELP TEACHERS WITH INSTRUCTIONAL DECISIONS.

MEETING TEACHER NEEDS -- NO SIMPLE SOLUTIONS

This discussion thus far might tend to lead some to think that one logical conclusion to our problem is to use more criterion-referenced tests and fewer standardized, norm-referenced tests. Unfortunately the discussion about these two types of tests has become narrow and oversimplified. Conventional wisdom in the current debate over testing is that criterion-referenced tests should be used rather than norm-referenced tests in order to discourage comparisons between children, to prevent misuse of test results by teachers and to avoid common abuses in the release of standardized test data. We need criterion-referenced tests, to be sure, to assist teachers in diagnosing student needs, judging student progress and individual needs and prescribing classroom remedies — the very kinds of uses teachers are now, often wrongly, making of standardized norm-referenced test information.

But we also need standards — and setting standards often involves making comparisons among children. How, after all, can we set an appropriate mastery level for a criterion- referenced test unless we have a sense of what the average child can do? And, how can we get a sense of "average" without a certain amount of standardization? In other words, is it really possible to develop a fair criterion-referenced test without administering the test to representative samples of children and examining their performance?

While teachers may not be as immediately involved in these processes and as immediately appreciative of their value as they are of other activities associated with test construction, these processes are no less essential to a comprehensive, quality testing program. In other words, an emphasis on the kinds of demands coupled with usage that our studies turn up should not be read to mean that standards and comparisons are irrelevant to teachers and schools.

Teacher needs go beyond even these types of tests. In the surveys discussed here, teachers felt a need to use standardized testing data to measure educational 'growth' or "judge student progress" (see Table 2). Unfortunately, one of the acknowledged problems of norm-referenced, standardized achievement tests is that in addition to not telling us much about what the individual child knows, they also cannot tell us much about how he is progressing. But we need to look at children over time if we are to get an accurate picture of the effects of schooling. Teachers need this information for their work. We also need it so that we can know more about effective schooling. Longitudinal studies that follow the same children for a number of years are remarkably absent in the literature of research. The development and use of more criterion-referenced tests should help us with this problem as well.

But this is not enough either. To really satisfy the needs of teachers — to really get a well-rounded picture of students — we need more varied forms of assessment. Sheldon H. White takes up this problem in "social Implications of I.Q." an essay in the compendium published by the National Elementary School Principals THE MYTH OF MEASURABILITY. White's essential argument is for an expansion in the types of tests we use to measure more accurately the range of intellectual diversity:

> Our experience with schooling tells us that children show diverse patterns of giftedness and achievement. This is true within the simplest form of elementary school as a place to foster reading, writing and mathematics. The similarities and differences among children concerning these skills are only lightly portrayed by a linear arrangement of grade-point equivalent scores on a standardized achievement test....
>
> I believe we must imagine that the reform of intelligence testing can best be accomplished by the widespread adoption of plural tests of human mental abilities ... such things as verbal ability, spatial ability,

reasoning, numerical ability, idea fluency, mechanical knowledge and skill, and so forth ... the inventiveness and use of such a system of characterizing differences among children would have considerable social benefits. It would provide a larger magic circle, encompassing significantly more of the reality one encounters in schools. It would also provide a considerably richer mixture of science in the midst of magic.

In other words teachers and other educators need tests and better tests — different tests for different purposes — to fill a wide variety of needs.

There is one other rather controversial point that needs to be raised in discussing the needs of teachers for test information. It begins with looking at what constitutes a good relationship between test use and teaching. If what teachers want is more information about individual students, and if we can assume they want it to assist them in their teaching, we can also assume that the existence of tests that provide this information will influence how teachers teach and what they teach. In other words, they may end up teaching to the test, a thought which provokes great distress among educators generally. The notion that teaching to the test is a bad idea is part of the contemporary mythology surrounding tests that deserves further examination.

In a very clever essay called "There Ought to Be a Law", Norman Frederiksen of the Educationall Testing Service takes a close look at this issue. Frederiksen tells a story of how a shift from paper and pencil multiple choice tests to tests that required students to perform tasks related to the operation of naval guns ultimately changed the way teaching was done in navy service schools. He notes that the change came about not because of any effort that was made to change the curriculum or teacher behavior. Improved student achievement and changes in teaching style were the direct and simple result of a change in the tests used. Frederiksen concluded:

> The moral is clear: It *is* possible to influence teaching and learning by changing the tests of achievement. It is also clear that those who make the tests have a great responsibility to produce tests that influence teachers to teach, and students to learn, the knowledge and skills that truly reflect ... objectives ...

Frederiksen goes on to discuss his own effort to develop such tests — tests that would seem to address the needs Sheldon White refers to, as well as the needs of the teachers who have answered the surveys discussed here. His tests are aimed at finding out about the psychological processes involved in problem-solving. Their titles are such things as "Formulating Hypotheses," "Evaluating Proposals," "Solving Methodological Problems," and "Measuring Constructs." Frederiksen's thinking and his work have led him to redefine tests: "A test is any standardized procedure for eliciting the kind of behavior we want to observe and measure. I mean the behavior we *really* want to measure, not merely something related to it." Actually, this definition of tests has been implicit in the ways teachers have used tests up until now. The problem has been that the tests have been inadequate to the task.

But if we really had the range of tests we needed — tests to measure a wide variety of behaviors and the learning skills these behaviors demonstrate, tests could help us refine the science of teaching in innumerable ways. In fact, Frederiksen ends up with a revolutionary conclusion:

> ....I have argued that it is possible to make tests that reflect instructional objectives more accurately than do conventional tests and that such tests influence the behavior of teachers and students in ways that enhance learning. If I am correct, it would seem sensible to use tests for teaching, not just for evaluation. Forms of a test could be constructed in such numbers and variety that they could be used regularly for homework or classroom drill. Students could cram and teachers could coach as much as they pleased. The cost of the tests would be justified by their value for instructional purposes.

If those who are now attacking tests could devote just a little attention to developing new tests and to helping teachers use both the new and the old more appropriately, education would gain much more than it is getting from the onslaught against standardized testing.

NOTES

(1). Jennie P. Yeh, "Test Use in Schools," Center for the Study of Evaluation, University of California, Los Angeles, June, 1978, page 28.

(2). Yeh, page 32.

TABLE 1

Use of Standardized Test Results in Instructional Planning

	Importance (4 = very important) Mean	S.D.
Student placement/grouping	2.7	1.12
Diagnosis of Individual Needs	2.8	1.11
Determining class needs	2.5	1.05
Judging student progress	2.5	1.04
Modification of your course content	2.3	0.96
Evaluation of your instructional program	2.4	1.04

TABLE 2

Use of Standardized Test Results in Instructional planning

by Grade Level

Importance (4 = very important)

	Primary (n=67)		Secondary (n=58)		Primary & Secondary (n=12)		No Grade Stated (n=5)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Student placement/ grouping	2.84	0.96	2.36	1.22	3.17	1.03	3.0	1.41
Diagnosis of individual needs	2.82	1.13	2.66	1.13	3.08	1.00	3.0	1.00
Determining class needs	2.59	0.99	2.47	1.16	2.42	1.08	2.60	0.55
Judging student progress	2.67	0.96	2.31	1.11	2.17	1.03	3.40	0.55
Modification of course content	2.45	0.96	2.16	1.00	2.42	0.90	2.60	0.54
Evaluation of your instructional program	2.51	1.05	2.16	1.06	2.67	0.89	2.80	0.84

TABLE 3

Influence of Formal Training in Tests and Measurement on Use of Standardized Test Results in Instructional Planning

	No Formal Training (n=14)		College Courses Only (n=60)		College Courses and Inservice Training (n=62)	
Student placement/ grouping	3.1	(0.86)	2.6	(1.17)	2.6	(1.11)
Diagnosis of individual needs	3.0	(1.04)	2.9	(1.13)	2.7	(1.11)
Determining class needs	2.7	(1.14)	2.6	(1.05)	2.5	(1.04)
Judging student progress	2.4	(1.16)	2.4	(1.08)	2.6	(0.99)
Modification of your course content	2.7	(0.91)	2.3	(0.99)	2.3	(0.95)
Evaluation of your instructional program	2.7	(1.27)	2.3	(1.07)	2.4	(0.96)

Table 4

Main Problems which Inhibit the Usefulness of Standardized Tests for Teachers

		n	%
h)	Results do not provide prescriptive information, e.g., guidance as to what materials, instructional activities are needed.	98	64
g)	Results do not provide an adequate profile of student strengths and weaknesses.	83	54
i)	Results are returned too late to be useful, or are not returned to teachers.	83	54
c)	Test content does not match my curriculum.	73	48
d)	Test materials are inappropriate and/or biased for at least some of my students.	71	46
e)	Comparison groups (norms) provided by the tests are not meaningful.	47	31
j)	Results are not reported in a form that facilitates interpretation.	46	31
k)	Results do not give me any new information about my students.	46	31
a)	Tests are given at the wrong time of year. A better time would be———————.	41	27
b)	Tests take too long to administer.	32	21
f)	Technical quality of tests is inadequate.	28	18

Table 5

Teachers' Perceptions of Information a Perfect Test Would Provide

Prescriptive information for each pupil	-40
Student's strengths and weaknesses	-32
Reasoning powers (analyzing, problem solving, etc...)	-19
Mastery of skills	-12
English/Language (grammar, vocabulary, spelling)	-11
Reading	-10
Grouping students by scores	- 9
Math	- 7
Writing skills	- 5
Comprehension ability	- 5
Does not exist	- 5
Provide information on curriculum taught	- 4
Personality - emotional (maturity)	- 5
Socioeconomic background	- 2
Potential ability	- 4
Verbal skills - ability to communicate	- 4
Strong and weak learning channels (i.e., visual vs. auditory)	- 4
Scores in relation to other areas, districts	- 3
Artistic ability and creativity	- 3
Overall factual knowledge	- 3
Motivation	- 3
Interests	- 3
Social knowledge	- 3
Self and environmental awareness	- 3
Teacher involvement needed	- 3
Ability to learn	- 2
Scores should reach teacher	- 2
Retention abilities	- 1
Physical	- 1
Why a student does or does not want to learn	- 1
Standardized testing is Big Business' profits	- 1
Concentration level	- 1
Effectiveness of teacher's instruction	- 1
Tests must be more complete	- 1
Tests should not confuse students	- 1
Should provide unbiased results	- 1
Do the children have emotional learning blocks	- 1
Leadership abilities	- 1
Learning growth (pre and post tests)	- 1
Ability for later employment	- 1
Chart - graph - map interpretation abilities	- 1

Table 6

Percents of Teachers Making Various Uses of Standardized Achievement Test Results in Their Classrooms

Personally use standardized achievement tests results for:	Total sample	Grades Combined Group 1	Grades Combined Group 2	Grades Combined Group 3	Groups Combined Grade K - 4	Groups Combined Grade 5 - 8	Groups Combined Grade 9 -12	Percent of Omits*
Individual student evaluation	65	63	60	80	65	68	55	7-11
Diagnosing strengths & weaknesses	74	74	70	84	77	76	63	6-9
Class evaluation	45	44	40	59	49	45	30	13-20
Instructional planning	52	51	51	58	52	56	42	10-16
Evaluation of teaching methods	37	36	36	44	40	37	29	15-20
Reporting to parents	42	41	40	54	44	46	28	13-20
Reporting to students	24	22	24	33	15	34	29	17-22
Measuring "growth"	66	67	61	77	71	66	43	8-18

*Percent of teachers in the various sub-samples who omitted this question.

Table 7

Percents of Teachers Who Consider Standardized Achievement Test Results Useful for Various Purposes

Standardized test results are useful to:	Total Sample	Grades Combined			Groups Combined Grades		
		Group 1	Group 2	Group 3	K-4	5-8	9-12
report to newspapers	10	10	10	11	8	11	16
report to boards of education	52	53	51	54	46	56	62
report to parents	67	66	64	78	63	70	70
report progress to students	56	55	56	63	44	66	71
measure educational status of individuals	61	61	60	67	58	64	65
measure educational "growth" of individuals	77	79	73	83	77	78	76
screen special education students	56	57	52	67	51	59	65
help plan instruction for individuals	63	62	61	70	61	68	59
help plan instruction for class groups	65	65	61	72	65	67	57
detect system-wide general strengths/weaknesses	75	76	72	81	73	77	79
help evaluate teaching procedures or methods	34	34	32	44	36	35	30
help evaluate instructional materials	41	39	43	46	41	42	39
help evaluate teacher performance	21	21	17	30	19	23	19
compare students with a national peer groups	58	60	53	63	54	59	69
compare classes in a school	30	28	29	36	26	32	36
compare schools within a system	36	33	37	49	33	38	41
compare a system with systems across the country	56	58	54	59	52	58	65

*Across questions and sub-groups, 5-12% of the teachers omitted particular question.

PART II

ASSESSMENT AND INSTRUCTION: WHAT MIGHT BE

In Part I, we discussed some of the main assumptions underlying conventional educational tests, and we examined the role these tests have played in the instructional process. As we argued in Part I, conventional tests have not provided much information helpful to teachers in the practice of teaching. One reason for the limited instructional value of conventional tests, we believe, is that the tests are based on a mistaken view of the relationship between assessment, teaching, and learning.

To understand the role of assessment in instruction, we believe it is necessary to begin by focusing on the ways in which teachers, in the day to day practice of teaching, form judgments about what their students have learned. Teaching, we argue, is an ongoing process of inquiry, in which teachers continuously draw inferences about what is going on in the minds of their students.

Conventional testing is generally conceived as something which either precedes or follows instruction — not as something which has instructional value in itself. But if the view we have taken is correct, assessment materials should be conceived as ways of expanding on the inquiry process already inherent in teaching. From this perspective, there should be little distinction betwen instructional materials and assessment materials.

We develop this view of the role of assessment in instruction in some detail in the two Chapters that follow. Philip Jackson, Professor of Education at the University of Chicago, examines the routine methods teachers rely on to make judgments about what their children know. Jackson identifies four common approaches teachers employ to draw conclusions about students' thought processes, ranging from informal observation to formal questioning and testing.

Each of these four ways of coming to understand students' thought, Jackson argues, is fallible, and taken together, the four methods cannot eliminate entirely the fundamental uncertainties involved in making judgments about students' cognitive skills. Furthermore, Jackson concludes, the act of using formal questions to test student knowledge can at times be disruptive of the teaching and learning process. Asking students continuously to demonstrate what they know can betray a lack of trust in student's autonomous capacity to learn.

We believe, then, that assessment materials for the purpose of instruction should not be viewed as something to be employed once insruction is complete. Instead, assessment materials should be viewed as materials much like regular classroom exercises or games — but designed to reveal strengths, weaknesses, and appropriate pathways through the curriculum. for individual students.

In the following chapter, David Hawkins explores some of the implications of this view of assessment, teaching, and learning. He begins by arguing that learning can be misrepresented in two seemingly opposing ways — as a process of transmission or shaping, and as a process of autonomous development. To understand the role of assessment in instruction, he goes on, both views need to be combined.

Hawkins argues that children are active model builders. They learn in the process of completing games, puzzles, and tasks. Learning is an activity in which the learner abstracts information from the world by selectively interacting with it, and many pathways are generally possible.

At the same time, learning depends on teacher guidance and direction. By focusing student attention on particular elements of a task, a teachers can increase the likelihood that the task will elicit critical skills and capabilities. By raising questions, a teacher can uncover hidden connections and deepen the quality of student discoveries. By assessing student interests, strengths, and weaknesses, a teacher can select appropriate curriculum materials and tasks.

Thus, Hawkins concludes, teachins is a dual process. The art of teaching involves both devising a curriculum and helping students find pathways through it. It involves both laying out tasks for students to complete and asking students to reflect on how they completed them. Assessment materials and instructional materials, then, are essentially similar. Tasks that encourage learning also provide information about the learning that has occurred.

THE UNCERTAINTIES OF TEACHING

Philip W. Jackson
University of Chicago

"A teacher affects eternity," Henry Adams once wrote, "he never can tell where his influence stops." That celebrated quotation, a mere twenty syllables in all, must surely come close to being the perfect tribute to the teaching profession. For what nobler thought could there be than the one expressed in its first four words and what truer fact than that contained in the remaining eight? "A teacher affects eternity; he never can tell where his influence stops." Inspirational, accurate, concise. A combination hard to beat. Small wonder, then, that Adam's verbal pat on the back, penned more than seventy years ago, retaines its appeal to this day.

Yet, however fine those twelve well-chosen words may be for chiseling in granite over the portals of schools or on the headstones of dear departed teachers, they leave much to be desired when read as commentary on the really troublesome uncertainties connected with the act of teaching. Adams never meant them to be read that way, of course. He obviously was more intent on paying respect to teachers than on being either descriptive or analytic about the details of their work. But questions about the more mundane and worrisome aspects of the ignorance from which teachers sometimes suffer are not long in surfacing once we have been stimulated to think about the more flattering forms of the unknowns they confront.

The mental process that guides our thinking about such matters seems to work a bit like gravity, at least it does for me. Just the way most things hurled into the air are pulled back to earth, so do my thoughts return to the here and now after a skyward leap of the imagination. And the more commonplace the topic, the faster, it seems, is the return. Teaching, being quite an ordinary activity, does not allow my ruminations to soar upward for long. After only a few seconds of wondering about the farthest reach of a teacher's influence I find myself asking questions like: What about the minute-by-minute influence teachers have on the pupils seated before their very eyes, an arm's length or so away? How much do they know about that?

"Much less, sometimes, than they would like to know" has got to be the only proper answer to such questions. For what teacher has not wondered from time to time whether this or that student really understood a particular point or whether the class as a whole was following the line of an argument or had grasped the moral of a tale? And who among us has had all such questions answered to his or her satisfaction? Surely the answer is: none.

So if we think of a person's influence as extending forward in time and space, as Adams's observation compels, it is not simply that the teacher cannot tell where his _stops_. In all likelihood he also cannot tell for sure where it _starts_ and from time to time he may have serious misgivings about how it is progressing between start and finish. Moreover, the latter forms of uncertainty can be quite unsettling, far more so as a rule than might any speculation about long term influence, for they bear directly upon such matters as the teacher's day-to-day sense of accomplishment and the public's confidence in the work of the schools.

As a teacher I may never live to discover that what I said one day in class has altered the course of human history a mite, and it is a pity that such good news is unlikely to reach me. But if I go home at the end of each day with serious doubts about whether anything I did or said had any effect whatsoever on anyone, I've got serious troubles, no matter what my future rewards might turn out to be. The public too might thank me and my teaching colleagues some day as it comes to realize what a powerful force for the good we have been. But if tomorrow it begins to suspect that our students are not learning what they are supposed to learn in our classes, the status of the entire teaching enterprise is in jeopardy.

The possibility of such deeply troublesome uncertainties arising among teachers or within the public at large does not make them a certainty of course. Indeed, they may never arise at all. No one, certainly, would wish them to. But the fact that we can even imagine them occurring and can do so with ease says something about teaching that we would do well to ponder, particularly if we are keen on preventing such unpleasant possibilities from happening.

A part of what it says has been stated implicitly already and is simple enough to be almost self-evident. It is that teachers may sometimes have a hard time proving their worth, even to themselves. Why this should be so is also easy to understand, deriving as it does from the obvious fact that teaching, unlike masonry or brain surgery or auto mechanics or even garbage collecting, has no visible product, no concrete physical object to make or repair or call its own. Consequently, unlike workers in the forenamed occupations and in the scores of others that could be added to such a list, when a teacher's work is finished he or she is without anything tangible to hold up as the fruit of his or her labor. No sturdy brick wall, no tumor-free brain, no smoothly purring engine, not even a clean back alley to point to with pride as evidence of a job well done.

Indeed, the very question of when the teacher' job is done, forget whether well or poorly, is itself problematic much of the time and must be established by agreeing in advance upon some fairly arbitrary cutoff point, a time to call it quits, such as a specified date on the calendar or a set number of instructional sessions. Moreover, what is true of the termination of instruction is equally true of resting points along the way. Even the decision to end a single lesson is more often determined by what the clock on the wall says than by any judgment of pedagogical accomplishment.

In this feature of their work, this absence of a tangible product whose gradual transformation yields a clearcut criterion of progress, teachers obviously are not alone. They are joined in this regard by ministers, priests, rabbis, therapists, performing artists, ambassadors of good will of all varieties — from office receptionists to public relation specialists — and countless other workers whose chief concern is with how some special group of people think and feel about things. At the close of the day, figuratively speaking, all these good people, teachers numerically prominent among them, wind up empty-handed.

Nor can it be said that teachers suffer more from this condition than do others who face if. There is no reason to believe that the psyches of pedagogues are any more or less sensitive to discomfort than are those of their fellow mortals who face a similar plight. Consequently, we might expect self-doubt and other forms of personal misgivings to plague teachers no more than anyone else whose labors yield little in the way of visible proof of accomplishment.

At the same time, granted that the general condition of periodic uncertainty occasioned by the absence of a tangible "product" is widely shared by many occupations and, in all likelihood, is equally troublesome to each, it is also highly likely that each occupation so burdened experiences and copes with this state of affairs somewhat differently. We might expect this to be if for no other reason than that the overall circumstances of each form of work — its mission, techniques, physical setting, and so forth — are sufficiently unique to set it apart from others. Why not, then, the uncertainties each face? Perhaps these too are uniquely defined for each occupation. An exploration of that possibility sets the agenda for what follows, which is to consider in some detail how the uncertaintes of teaching are commonly and perhaps uniquely thought about and dealt with. When such a close look is taken what emerges is a view of teaching that is at once familiar and strange.

What puzzles teachers most? What is characteristically problematic for them? How do they think about the uncertainties they confront? There are many ways of framing the opening question of such an investigation, but none has a definitive answer, for the circumstances of teaching and the personal characteristics of individual teachers vary enormously and change over time, as do the broad features of the profession as a whole. Consequently, what is puzzling for one teacher may not be for another and what teachers of today look upon as problematic may have been taken for granted or never even examined by their predecessors a few generations back. Yet despite these situational, personal, and historical variations, there are similarities and continuities as well in the way teachers characteristically view their work. With respect to the brace of questions used as openers, the answers with the broadest applicability across different settings and different times would surely contain some reference to two closely allied perspectives on the teacher's task. One of these is philosophical in orientation; the other, psychological.

Philosophically speaking, all teachers might be said to be puzzled chiefly about epistemological matters. That is, one of the most common ways of talking about the goal of teaching is to describe it as having to do with knowledge and its transmission. Accordingly, when it comes to the question of what worries teachers most we might reasonably expect that the answer would have something to do with the status of some specific bit of knowledge, be it a skill, a propositional statement, a logical construction, or what have you. And we hardly need conduct an empirical investigation to affirm that expectation. Anyone either who is or has been a teacher or who has been around teachers for any length of time (and the latter category must include almost everyone) would surely agree that teachers seem to spend a lot of time worrying about that most ancient of all dichotomies: THE KNOWN AND THE UNKNOWN.

But this recurrent concern, which I have christened with the adjective epistemological, could as easily be called psychological as well. Though teachers may be accurately described as being principally concerned with the status of some body of knowledge, they are not concerned with it in the same way as would be a person studying that knowledge on his or her own, nor as someone seeking to add to that knowledge, nor yet as someone chiefly interested in the principles or conditions by which knowledge in general comes to be established as might, say, a cognitive psychologist or even someone who called himself a professional epistemologist.

For one thing, teachers are chiefly interested in the status of other people's knowledge, as compared with their own. But that does not set them apart, of course, for there are many people who are interested in a professional way in what others know or do not know. (Public pollsters and spies come immediately to mind.)

What distinguishes the epistemological puzzlement of teachers, if I may stick with such a fancy tag for the worries under discussion, is that it focuses on knowledge that is or is not lodged, so to speak, in the minds of an identifiable (and usually a clearly identified) group of people, called students, and on knowledge for whose transmittal the teacher is either partially or wholly responsible. This means, first, that of all the uncertainties facing a teacher some of the most bothersome take the form of questions about WHAT IS GOING ON AT THIS INSTANT INSIDE THE HEADS OR MINDS OF THE PERSON OR PERSONS BEING TAUGHT. Do they understand? Are they following me? Has he grasped the point? A parallel set of questions fills the pedagogical mind when instruction has ceased. Did they understand? Have they now achieved mastery? And so forth.

It means, second, that the teacher's answers to such questions, even his guesses as to what the answers might be, have an important bearing not only on what his next move will be as a teacher but also on his notion of how successfully he has performed his work.

This is not to say that no one but a teacher raises questions about whether another person does or does not understand whatever the questioner has been trying to communicate. Such queries are commonplace in human affairs. They occur each time someone says to someone else "Do you understand?" or something equivalent. Usually, however, the "messages" whose acknowledged receipt is being sought in such exchanges are situationally specific in content and, therefore, do not qualify as knowledge that is generalizable to many situations the way the contents of a teacher's lesson purport to do. When what is being communicated does have such a generalizable quality, the exchange is decidely "teacherish" in tone, no matter where it occurs or whether any of the participants think of themselves as either teachers or students.

Having said this much about the epistemological and psychological focus of a teacher's concerns, we are ready to ask how he or she typically goes about responding to them. What, in other words, does the teachere do to answer the many questions that crop up during the process of teaching? Once again, we need not initiate a tedious empirical investigation to obtain at least rough and ready answers to this, our second order, question. Given the familiarity of teaching to most of us, all we need do is to picture in our mind's eye a typical classroom teacher at work. By so doing, most of us can easily "see" what an answer to our question — at least in gross terms — would have to contain. By this easy exercise of our imagination we can, as it were, envision the major ways in which actual teachers may be seen to go about the business of finding out what is going on inside the heads of their students. According to my own count, there are four such strategies. In real life not every teacher may be found to use them all, and some teachers (such as those on television) may use none at all, but each is common enough to be familiar to most of us. The first three involve actions that take place while teaching is going on. The fourth occurs only after teaching has ceased, has been temporarily halted, or has not yet begun.

The least formal and the least intrusive of these four ways of investigating what is happening in classrooms is the common one of looking around the room for signs of the students having difficulty with what is being taught. This form of visual monitoring is most readily observable when the teacher is delivering a lecture or conducting a discussion, though it can sometimes be seen to occur during the supervision of seat-work and study periods. What the teacher is looking for on such occasions are those spontaneous indicators of understanding and interest or the lack thereof that can be "read", so to speak, from the looks on students' faces and the postures they adopt. These include nods of assent, smiles, frowns, furrowed brows, head scratching, fidgeting, droopy eyes, and much else that makes up the "vocabulary" of what is sometimes spoken of these days as "body language."

A standard way of talking about this kind of visual search is to say that the teacher is trying to find out whether or not the students are _with him_ or whether they are following him in their understanding. If the judgment is that they are not, they are sometimes spoken of as being _lost_ or _out of it_, a condition calling for some kind of remedial action. Finally, though the chief purpose of the teacher's visual scan may be to seek information about how things are going ("things" referring principally to the students' understanding of the material being taught), the act itself is often perceived by students to be a kind of warning signal, reminding them to remain attentive and alert. Thus, the procedure itself helps to bring about the conditions that are the object of the search.

The second of the four techniques is not as easily observable as the one just described, though it is hardly less common. Its lesser visibility derives from the fact that it has more to do with the establishment of a classroom procedure than with any readily identifiable movement or action on the part of the teacher. Basically, the procedure is designed to encourage students to volunteer information about the status of their understanding of the material being taught. Usually this encouragement takes the form of an invitation to interrupt the teacher or the classroom proceedings whenever there is a failure to comprehend what is being said or done, though the formality of actually inviting distress signals of this sort is often unnecessary. Many students volunteer the information without being asked. (Indeed, sometimes the interruptions come so thick and fast that the teacher is obliged to slow them up or stop them completely, usually by requesting that such questions be held until the end of the class or until there occurs a natural break in the session.) In essence, then, this strategy amounts to arranging conditions so that students will call for help when they are in trouble, thus signalling a breakdown in comprehension or understanding.

A third common way of finding out whether or not students understand what is being taught is to ask them directly while teaching is underway. Such questioning takes many forms, most of which can be arranged along a continuum of specificity that refers to both the content of the question and the person or persons to whom it is addressed. At one extreme are those queries addressed to no one in particular and calling for little more than a nod of the head or a showing of hands. These are often one-word questions, such as "Understand?" or

"OK?" or "Right?" Some teachers use them so habitually that it is doubtful that they are even aware of doing so.

At the other extreme, and much more interesting from the standpoint of understanding what school is all about, are questions to individual students, asking them to display their knowledge in some detail or to perform a particular skill for the teacher's inspection. These targeted queries leave no choice but to respond in one fashion or another, thus revealing knowledge or ignorance for the teacher and all others present to observe. Indeed, an old fashioned way of dealing with the answers given was to grade and record them on the spot, a procedure that was part of what used to be called the recitation method.

Fourth and finally come the most formal of all teaching procedures aimed at finding out what students have learned. As anyone who has ever been to school must by now have guessed, these comprise tests, quizzes, exams, and related activities that typically occur during lulls in teaching or after it has ceased completely. In addition to the ubiquitous paper-and-pencil tests they include term papers, oral examinations, project reports, recitals, and other means of allowing or requiring students to display their newly acquired knowledge and skills. Beyond occurring outside of teaching, so to speak, these forms of questioning (for that, in one sense, is what they all are) have an official quality and an air of finality about them that customarily are lacking in the less formal methods that have been described. This is so because their results commonly serve as the chief, if not the sole, basis for assigning course grades.

Here then, if my exercise of imagining a typical teacher in action has yielded an accurate portrayal of reality, are the four most common ways employed by teachers to quell whatever uncertainties might arise in their minds about what is happening or has happened in the minds of their students. There may be other common ways as well, but none suggests itself to me. Consequently, I offer these four as the classic procedures by which teachers cope with the unknowns that beset them.

How successful these procedures turn out to be will depend, of course, on the skill and consistency with which each is employed. Some teachers are doubtlessly more skillful than are others in their use and some teaching situations lend themselves more easily to their application than do others. Such differences aside, however, it can be said of all four that none is foolproof and that each has special shortcomings limiting its usefulness. Some of these limitations are widely recognized and understood; others, seem not to be.

It is well-known, for example, that the outward signs of inner attentiveness and understanding can be faked. Thus, by looking around the classroom and relying on visual cues alone the teacher may think that everyone is following the discussion or whatever, whereas many may not be. Conversely, the student who appears to be dozing off in the far corner of the room may actually be the most attentive of them all. Such are the ambiguities that plague the application of the most effortless of the four methods.

We know too that calling for students to signal their own difficulties has built-in drawbacks. Though the teacher may do everything in his or her power to create a non-threatening atmosphere, one in which students feel free to say what is on their minds, and to confess to troubles when they arise, not everyone, even in the most comfortable environment, is willing or able to take advantage of such an opportunity. Consequently, no matter how hard the teacher might try to have it otherwise, there will always remain the nagging worry that some students are having difficulties in understanding but are not saying so.

Turning from these two more or less passive strategies to the two more active ones, those involving questions the teacher puts directly to one or more students, we find the fallibility of the information they provide to be somewhat different in quality but no less troublesome. In fact, the use of these direct probes and even the threat of their use introduces into the teaching encounter an element of social tension and an unusual quality that serves to set teaching apart from other forms of human activity. But before examining these more subtle features of the questioning process as it occurs in classrooms, it is well to take note of some of its more obvious limitations. Only by so doing can we begin to understand why formal evaluative procedures, such as tests and quizzes, are not more widely used in schools than they are.

Those questions the teacher asks of the class in general — queries like "Understand?" or "Is that clear?" — are so obviously open to false answers (or to no answer at all) that little more need be said about them. It is worth noting, however, that signalled comprehension or understanding can be false in two ways. It may be that the student who nods his head when the teacher asks: "Understand?" is aware that he lacks understanding but wishes to hide that fact from the teacher. But it may also be that he thinks he understands, but truly does not. Thus the unreliability of the information yielded by this form of questioning has two potential sources.

Questions have content and that are directed at particular students may not leave the teacher guessing whether the questioned student does or does not understand what is being taught (though poorly phrased questions can leave much in doubt) but they have drawbacks as well. The most obvious of these is that an unsuccessful or incorrect reply is commonly a source of embarrassment to the person giving it. It can also be a socially disruptive event for the class as a whole. Consequently, a standard practice among teachers seeking to reduce the likelihood of such "wrong" answers is to pose questions to the class as a whole and then seek volunteers to answer them. This procedure is obviously designed to avoid the embarrassment of calling on someone who must then confess ignorance. But the ploy is by no means foolproof. The degree of underestanding signalled by the waving hands of volunteers can be either more or less than it appears, as every teacher knows.

Added to the threat of embarrassment associated with direct questions from the teacher while class is in session are economic constraints as well. Such questioning obviously takes time, which commonly means time taken away from direct instruction. Moreover, once a question has been asked and answered in public its pedagogical usefulness is spent. (Teachers can and do follow up successful answers with queries like "How many agree with Sarah?", but the reliability of the information received in reply is generally not much greater than when the teacher asks, "Understand?") So in addition to using up precious class time such direct questions have to be employed judiciously for they commonly are not reuseable.

Incidentally, a common pedagogical practice that avoids many of the pitfalls and limitations being discussed is to avoid questions that have correct or incorrect answers and concentrate instead on eliciting student opinion or attitude. This tactic obviously eases the social strain and makes it possible for the same question to be addressed to more than one student. "After all," a teacher using this technique might point out, "everyone is entitled to his or her opinion." The trouble, of course, is that not all curricular content lends itself to such a non-threatening sharing of individual viewpoints. Indeed, critics of this pedagogical strategy might call it an avoidance of the teacher's responsibility for the advancement of his students' knowledge. Exchanging opinions might be fun, the criticism might concede, but seldom does it promote any true intellectual gains.

Turning from the kind of questioning that goes on while class is in session to that comprising paper-and-pencil tests, term papers, and the like, we face many of the same limitations that already have been discussed and some new ones as well. Tests, like the directed questions teachers raise in class, are threatening to many students. they are costly in time and energy to construct, administer, and score. Because of such costs they almost invariably are limited to a sampling of the questions that could be asked or even of the ones the teacher would like to ask, and frequently a very small sampling at that.

From the standpoint of its usefulness to the teacher himself, the information gathered through such formal procedures is seldom of much direct value, for it typically arrives too late to be of help to the teacher in modifying what goes on in the classroom. Assessment procedures that are part of some of the newer schemes for individualizing instruction (e.g., IGE, IPI, etc.) may be exceptions to this general rule, but by and large the rule stands: Tests are relatively ineffectual means of clearing up whatever uncertainties teachers may have about how well or how poorly they are doing their job. Methods of evaluating students that are even further removed from a direct display of knowledge gained through instruction (such as term papers, projects, and the like) may provide the teacher with useful information about many aspects of a student's performance, but, again, they are unlikely to reduce any of the uncertainty that might exist concerning the effectiveness of the teacher's own actions.

Here, then, are several of the more obvious drawbacks associated with the four most common ways teachers go about the tricky business of trying to find out if the material they are teaching is getting across to students. The purpose of highlighting the fallibility and limitation of each method is not to suggest that teachers should use any of them less than they do. Rather, it is to begin to explain why some of them, paticularly the more formal and direct methods of questioning, are not used more frequently than they are. Moreover, with respect to the latter procedures, two further considerations need be added to those already mentioned. Both have to do with the somewhat peculiar nature of the questions teachers ask.

Normally when people ask questions they not only expect answers, they need them. That is, they are seeking the information requested for its own sake. (There are, of course, exceptions to this rule, such as rhetorical questions and those "polite" inquiries to which a standard response is usually given — e.g., "How are you?") Indeed, in everyday affairs if we are given cause to believe that the person asking a question already possesses the information being sought we would legitimately begin to wonder why he or she bothered to ask. Were they simply teasing? Were they trying to catch us in a lie? Were they seeking a confession? Whatever the answer, we would be reasonably confident that something was fishy about such a state of affairs.

Consider, however, the condition that obtains when a teacher calls upon a student to display a piece of acquired knowledge or skill. The questioner in this instance already possesses the information requested. What he does not possess, of course, is the knowledge of whether the student being questioned can accurately or faithfully produce the known answer. So the teacher's real interest is not in the content of the answer _per se_, as it is in most other everyday situations, but rather in the student's ability or lack thereof to deliver the expected reply.

This is not to say that teachers commonly disguise their true intent nor that they could do so successfully should they try. Except perhaps at the very lowest levels of schooling — kindergarten or thereabouts — most students know full well that when a teacher asks a question it is commonly to find out whether they (the students) know or can do something and is not a search for the about-to-be-displayed knowledge _per se_. Teachers rarely if ever go out of their way to hide this fact. Nor is there any reason for them to do so. It is widely understood and accepted by students and teachers alike that an integral part of the teacher's task is to become reasonably certain that a particular piece of knowledge or skill has been acquired. What better way to accomplish that goal than the kind of direct questioning being described here?

At the same time, even though it may be perfectly legitimate for teachers to ask questions as they do, and quite understandable as well, there is something about the circumstances and the format of the inquiry that injects a note of artificiality into classroom proceedings. It's as though the teacher were somehow acting or pretending or even playing with students rather than responding to them forthrightly and openly. For even if it is the teacher's legitimate duty to try to find out whether or not a student knows something, the process itself often has a kind of cat and mouse quality about it that is rarely present when people ask questions in out-of-school settings. The teacher, if he or she wanted to, could as easily _give_ the student the answer as request it. This must mean not simply that the teacher possesses the information being sought, as has already been acknowledged, but also that he or she prefers, for the time being, to keep it hidden. Is there not an element of teasing in such a posture? Might not a perfectly natural reply to a teacher's query be: "Awww, _you_ know"?

And beyond the playful quality lies something even more disquieting to contemplate. For, come to think of it, shouldn't the teacher often be in a position to _know_ whether or not the student knows something even without asking? After all, it is the teacher's job to see to it that the knowledge gets delivered, so to speak. Indeed, he or she often delivers it in person. What can it mean then for a teacher to ask a student if he knows or understands something that he has just recently been told? What are the sources of the doubts that might lead to such a question?

The first thing to say about them is that they are multiple. All kinds of mishaps may occur between the teacher's delivery of the knowledge or his recommendation that it be obtained from somewhere else (e.g., a textbook) and its safe deposit, so to speak, in the student's memory bank or neurological network or however one wishes to conceptualize its resting place within the person. The student may not have heard or seen what was said or done. He may have received the message but not comprehended its meaning. He may have understood something perfectly a short while back but now forgotten it. And so on.

Moreover, all these envisioned mishaps and more that could be named have a conceptual source that sets limits on our understanding of all that can go wrong. They are rooted metaphorically in the image of the student as some kind of container or vessel in which knowledge can be stored. Depending on whether knowledge is itself conceived of as being solid or liquid, the task of the teacher, within the terms of this metaphor, is to see that a sufficient quantity of this precious commodity is packed or poured into the students under his charge.

But there are other ways of conceptualizing the teaching-learning process beyond depicting it as a mechanical operation involving little more than filling the heads of students with a load of knowledge. Each of these alternative metaphors calls attention to additional difficulties that teachers might face. For example, if we think of knowledge as being like food that is digested, rather than as being like an object that retains its original form or shape inside its container, we can begin to envision the teacher as having a quite different set of worries, many of which add to the urgency of his questioning. Instead of wondering whether some nicely wrapped parcel of knowledge lies safe in the shelf, so to speak, somewhere within the student, he now begins to worry about whether it has arrived in one piece, how it matches the knowledge that was there before, how it gets used by its new owner, and so forth.

These alternate ways of imagining what goes on when teachers try to teach do little, if anything, to reduce the tension implicit in questions that call for a display of knowledge or understanding. Indeed, in some ways they may be said to increase it. That tension derives in part from the fact that the teacher's query all too often threatens to produce a rupture in the social relationship between teacher and student. The dynamics of this threat are revealed in the following vignette.

Suppose a gift of china dinnerware is sent as a wedding present to the home of a prospective bride. A few days later the gift-giver calls the home of the bride-to-be to see if the gift arrived safely. "Yes it did," is the answer. "I'd like to see for myself," the caller replies, "I'll drop by this evening."

What's so strange about that situation? Well, quite obviously, the odd part is that the giver of the gift does not trust the testimony of the bride-to-be. There is nothing peculiar about his calling to see if the gift arrived, true enough, but ordinarily we would expect his inquiry to cease once he has been told that the gift had reached its destination. His failure to do so is a serious breach of social etiquette.

Though teaching only remotely resembles gift-giving, an interpersonal relationship similar to the one in the situation described threatens to come into being when teachers insist on having students display in detail the knowledge they possess. The resemplance is particularly close, of course, when the teacher's direct question has been preceeded by a general query concerning the understanding of the material being taught. "Did the knowledge arrive?" asks the teacher. "Yes," nods the student. "Let me see," says the teacher. "What's the matter, don't you believe me?" asks the student. "Oh, sure I do," the teacher replies, "It's just that ...."

That what? Were the teacher pressed to give a frank answer to the student's query, one that he may have difficulty facing up to himself, I fear it would be that something resembling distrust does lie behind the demand for hard evidence of learning's having occurred and, much as we might wish it were otherwise, such suspicions often turn out to have been warranted. For the truth is that there are many reasons why peple might try to hide the fact that they do not know something, even people who are usually honest about most other things. Ignorance is often an embarrassing condition, no two ways about it. It is especially so in a classroom after the teacher in charge has made an effort, either direct or indirect, to assure that something has been learned. Under those circumstances the student who admits to not knowing what the teacher set out to teach has confessed to having failed in one way or another — failed to have listened, failed to have understood, failed to have done the assignment, or what have you. He may ultimately be excused or forgiven for his inability to respond satisfactorily, but its status as a failure remains.

Thus, it is not terribly surprising to find that many students will not voluntarily expose their ignorance and will even try to keep it hidden when others, such as a teacher, threaten to reveal it through direct questioning. So the suspicious attitude that lies behind the seemingly innocent query from the teacher is not the sign of a streak of paranoia in his personality. It is, instead, an understandable preparedness based on a realistic appraisal of human nature.

But the legitimacy of the teacher's suspicions does not make the act of putting them to rest any more comfortable for either party. It is awkward, to say the least, to have to check up on people and it is demeaning, if not downright insulting, to have to be checked up on. However much we might try to avert the discomfort connected with such a query (and many teachers seem to be quite skillful at removing the sting from their questioning) it is doubtful that the process can ever be totally painless.

To recognize this fact is not to argue for the abandonment of tests or the elimination of direct questions in class or anything of the sort. If teachers are to fulfill their professional responsibilities, they often have no choice but to insist that students display their newly acquired knowledge, or the lack thereof, no matter how painful or embarrassing such a disclosure turns out to be. At the same time, recognizing the threat of discomfort implicit in direct questions, tests, and the like, we can begin to understand why some teachers might hesitate to employ such procedures; why, in other words, they might prefer to live with the uncertainty of not knowing for sure whether their students have in fact learned what was taught. The costs of obtaining that information must be weighed against not only the discomfort it might bring to individual students but also the potential damage it might do to the social relationships involved. We may condemn the teacher who avoids at all costs the slightest threat to a warm and comfortable relationship between himself and his students, as we might the parent who never disciplines his child, but we can at least understand the motives that guide him along such a course of action.

Where, then, has this discussion of pedagogical uncertainties taken us so far? It has, I trust, underscored the central fact with which we began, which is that the process of teaching, viewed as knowledge transmission, is fraught with unknowns. In holding up for brief inspection what seems to be the four major ways in which teachers cope with this condition, it has also revealed some of the limitations of each of these strategies for finding out what is going on "inside the heads" of students. Some of those limitations have to do with the fallibility of the information each strategy yields; others with the costs — economic, psychological, and social — connected with its use. The upshot of this analysis may not be new, but it is important nonetheless. What it suggests is that in teaching as in most other complex activities, the path of reason is often forked. Just as it makes good sense for a teacher to want to know whether or not his students are learning what they should, so does it also make sense, and often equally good sense, for him to avoid the very kind of questioning that will yield the most reliable answers to his pedagogical inquisitiveness.

How teachers handle this tension between wanting to know what is being learned but not wanting to spend too much time and energy in finding out and, at the same time, not wishing to create an undue amount of social discomfort in the process, is partially an individual matter. Some teachers seem content to press such queries no further than what they can see with the naked eye, others insist on questioning almost every student at almost every turn. Some use quizzes and exams whenever the opportunity permits, others eschew formal tests completely.

But not all such variations are a matter of personal preference. It is also doubtlessly true that some curricular areas lend themselves to direct questioning more easily than do others. We know, for example, that mathematics and spelling are more adapted to paper and pencil tests than are, say, social studies or literature. Moreover, rudimentary levels of understanding are usually more easily revealed by direct questioning than are higher levels. Thus, we might expect to find a heavier use of such procedures in the earlier grades than in later ones.

Beyond such variations in the adaptiveness of curricular content to the strategy of direct questioning lie differences in the level of social concern aroused by the threat of people not knowing what they are supposed to know. In short, we worry more about whether some people are knowledgeable than we do about others. We seem to care more, for example, about whether a physician "knows his stuff" than we do about, say, a florist Consequently, we would expect teachers in a medical school to be somewhat more conscientious and demanding about asking questions and giving tests than we would teachers of floral design.

The overall level of such worries seems to change over time as well. Right now we appear to be in the midst of a period of heightened public interest in the outcomes of schooling, particularly at the secondary level and below. Consequently, we hear a lot of talk these days about such notions as educational accountability and minimal competency testing. How long the present trend will continue remains to be seen, but so long as such a mood prevails teachers are bound to feel additional pressure upon them to seek "hard" evidence of what is or is not being learned by their students.

An additional spur to the employment of direct quesions in the classroom, particularly formal tests, comes from the growth of the technology of test development and the associated emergence of the testing industry. These developments likely have a double effect on what teachers do to find out what their students know. On the one hand, teachers these days are better trained in the techniques of test construction than were their counterparts a generation or two ago. On the other hand, today's teachers also have access to a vast supply of commercial tests and workbooks that were not available in the past.

Furthermore, the development of mass testing programs that lie outside the realm of teacher decision-making (such as the SAT or the National Assessment of Educational Progress) doubtlessly heighten the overall desire of teachers to be sure that the material they are teaching is getting across.

Given the complexity of this mix of forces impinging on the teacher's decision to question or not to question, to test or not to test, about the only thing that can be said for sure about such decisions is that they probably are not as easy to make as they might first appear to be. Two groups in particular, it seems to me, tend consistently to underestimate the difficulty of the teacher's position in such matters. The first comprises the bulk of our so-called experts in the field of educational testing and evaluation. The second is made up of the majority of today's advocates of a let's-get-tough-with-students policy.

In addition to overlooking some of the psychological and social costs of questioning that have already been mentioned, both the testing experts and the citizens clamoring for greater accountability usually suffer from another kind of short sightedness as well, which is brought about by their almost exclusive reliance on the particular view of teaching that has been dominant in this essay. That view, as has been said several times, depicts teaching as essentially a process of transmitting knowledge.

Now there is nothing _wrong_ with this outlook on the teaching process, to be sure. Indeed, there seems to be a lot that is _right_ with it. The important question, however, is whether such a perspective affords a total view. In other words, is that all there is to teaching, the transmittal of knowledge?

Some people, like Mr. Gradgrind in Dicken's _Hard Times_, would certainly say yes. Indeed, even knowledge was too highfalutin a term for old Gradgrind. As a teacher all he wanted to get across were "Facts, children, facts!" A few flesh and blood teachers doubtlessly would echo the same sentiment today.

But the majority, I suspect, would be unhappy with such a narrow view. Even those teachers who are willing to accept as the central purpose of their work what I have called its epistemological character would probably insist that there is more to it than that. How they talk about the larger scope of their mission, whether they discuss it in terms of character development or moral education or aesthetic appreciation or social responsibility or whatever, matters less here than does the fact that none of these ways of talking is reducible to language that is strictly epistemological. All, in other words, refer to modes of experience and to psychological states that spread beyond the boundaries of knowledge _per se_ and that are not easily revealed, if at all, by questions from even the most skillful teacher or test-maker.

There are even times, it seems, when the most sensible thing for a teacher to do at the end of a lesson is to remain silent, or close to it. Elizabeth Hardwick, teacher and writer, describes one such occasion. "It's hard to say anything about a fine short story," she tells us, "I know from teaching that I would ask the class to read Chekhov and all I could think to say to them was, 'Isn't he wonderful!'" Most teachers have had similar moments of speechlessness in all probability. I know I have. At such times the question of whether some piece of knowledge is or is not lodged in somebody else's head seems like a silly thing to want to know. So too does the broader question of precisely what influence the teacher's actions have had. We can do little better on such occasions than to join with Henry Adams in his celebration of all the things that teachers will never know. Those uncertainties begin afresh with each new day of teaching and seem to have no end. Adams hit the nail on the head all right in what he had to say about the farthest reaches of the teacher's influence, but he could as easily have used the close at hand as his starting place. "Near and far," he might have said, "the teacher's lot remains the same — from here to eternity, uncertainties galore."

THE UNPREDICTABLE NATURE OF LEARNING

David Hawkins
University of Colorado

In this paper I wish to consider that aspect of educational assessment which is primarily of use to teachers in the exercise of their art. I shall be speaking mainly of the elementary school ages. In order to consider this aspect I shall however lay down certain general propositions about the process of education, of teaching and learning, and about the word "curriculum."

In a genetic sense education is a process which can be misrepresented in two apparently opposing ways, each of which catches something of the essence but each of which is incorrect if translated into practice and is inconsistent with the other. Some things are complicated enough to require at least *two* sentences to say them. And as in mathematics, two axioms taken together may generate a nest of theorems which would in no way follow from either of them alone.

The first of my axioms is that education — informal education first, formal education added — is the central process of culture *transmission*. By culture I mean everything which contributes to children's potential capacities to become competent functional members of their society — including all relevant aspects of knowledge, skill, character and commitment. The metaphor dominant in discussions of this aspect is that of the potter and the wheel, the metaphor of *shaping*. Human beings are in some measure plastic, and from birth are being instructed, molded, shaped. In culture transmission and culture evolution education takes the place of the genetic code and subsequent embryology. Child development, so considered, is the interaction of social nurture with embryology.

In narrow applications of this view the attempt can be made to assimilate the description of the process of education under the metaphor of standard engineering design. Our public education system, dealing with numbers of teachers in excess of a million, has evolved — over a very few generations — creating an institution which has in it, across the land, many dominant informities of practice, of daily and longer-term routine, of style and practice. This standardization brings with it, understandably, certain aspects of quality control relating to various levels of assessment and accountability.

In a simplistic account of engineering design two presuppositions are basic. One is the availability of uniform raw materials of known properties, the other is a system of rules or procedures for shaping and assembling these materials into a finished product. In reality, however, these assumptions are only approximated, and it is necessary, as part of the design itself, to monitor for non-uniformity, for choice among alternative rules, for chance deviations.

In bringing this point of view — at *some* levels of approximation a necessary one — to bear on the process of schooling one is forced to recognize a very considerable non-uniformity among children, among teachers and their practices, among schools and systems, curricula, etc. Among the many sorts of monitoring assessments which this situation invites is the constant assessment of children's progress along standardized curricular tracks. This may be the basis for routing children or youths among alternative tracks, or for evaluation of teachers and schools, etc. Such assessment itself requires some measures of standardization, as for example in the comparison of schools, systems, for making national comparisons, or across time. In recent decades a dominant response to this demand has been the creation of a wide variety of statistically standardized measures, almost invariably paper and pencil tests, and these tend to become implicit definitions of educationally desirable objectives. What is outstandingly obvious is that their results reflect a quite gross variance with respect to the erstwhile uniformities which the metaphor of engineering design has presupposed; much of this variance remains unaccounted for except in terms of conventional ideas such as "ability," etc.

I now turn to my other axiom. Human beings are by nature active model builders; their learning — from birth — is essentially an autonomous process in which their behavior (conduct) is being constantly modified by processes of assimilation, accomodation and equilibration (Piaget) which involves the mapping of environments and the planning of conduct, both processes taking place at levels of motivation and informational complexity which take account of motor-sensory input but which are not accounted for by external sensory input (including "reinforcement").

Such input is in part an independent variable, but in part is information elicited by the individual, in part dependent on his activity and discrimination. Those aspects of nurture and environment which are relatively independent of such elicitation will indeed have a directive influence on the models built, and support or discourage children's general model-building properties, which are by their nature cumulative or autocatalytic (intelligence).

In the course of such careers human beings are congenitally diverse in their model-building motivations and propensities. Beginning from an initial genetic diversity these differences become amplified in some essential respects, but also can be seen as alternative pathways along which common social characteristics of habit, language, of institutional accomodation — are or can be developed. What appears from the viewpoint of the first axiom to be non-uniformity is from the point of view of the second, more adequately described as what we call individuality.

From this point of view the readiness for learning is primarily a matter of individual development to date, of individual motivation. The metaphor of transmission becomes inappropriate; learning is primarily an activity of the learner, abstracted from information selected or elicited by the learner from primary subject matter, from the accessible world, through his selective interaction with it. In this activity the learner is an eolithic craftsman building structures — models — of his own, using what has been already assimilated, including frames of thought already stored from previous learning, with ends-in-view which are themselves framed in terms of prior experience

The role of teacher, seen in the light of each of these axioms in turn, and excluding the implications of the other, is a kind of stereotype. Under the first axiom the central role is that of instruction, leading students along a pre-determined pathway, on their part a step by step acquisition of skill and knowledge, shaped — informed — by the teacher as source, or -nowadays more typically- by the teacher as administrator of standardized sources — textbooks, workbooks, "packages."

Under the second axiom alone the role of the teacher is no longer primarily that of an instructor. The teacher becomes a guardian, a facilitator, a "support facility," organizer of a material ambiance in which children's model-building propensities will be supported, providing materials which they can shape in accordance with these propensities, each in his own way and according to his own readiness and momentary motivation. If there is educative direction in this provisioning it is indirect; if there is instruction it is instruction of demand, assistance in pursuit of an end set by the learner.

In a superb philosophical essay, still in print but seldom read with any due regard for its content, John Dewey(1) sets forth the dialectical development of these two axioms when they are firmly brought together. His first step is to set forth each of tese axioms — as I have called them — in such a bald and stereotyped form that they appear to contradict each other, not only in logic but in a whole stream of practical consequences which each seems to entail. These contradictions become the armamentarium of warring parties in a perennial debate, each charging the other as espousing ideas and practices which doom education to failure.

Dewey's second step is surely the right one; it is to say — in effect — that both axioms are correct, and that each, taken without regard to their joint implications, will in fact bring about the failure which it is accused of. Without accepting both axioms, in some suitably refined form, one simply cannot define the central problems of education.

Unless the classroom is both child-centered and subject-centered the basic conditions of educative success will not be met. The teacher's central role is that of bringing about a match between the child and the curriculum in an enriched environment. Such an environment entices children's curiosity and gives them wide access to subject matter. It leads them into the curriculum by selecting, reorganizing and embodying its content in that environment, thus "directing by indirection." Dewey was aware of the fact that there is a large multiplicity of pathways into the exploration and final mastery of any domain of elementary subject matter, and that it is only by the teachers' art that pathways can be found to match the propensities and talents of individual children, and sponsor the kinds of associated activities which will bring them, as a small society, to relive the intellectual and practical learning and invention of mankind. Dewey discusses at length the contrast between the standardized logical organization of subject matter (e.g. the textbook) and what he calls the psychological organization, that from which a teacher, knowing well the logical organization, will reconstruct accessible content from it to maximize access and commitment from diverse individuals and groups of learners.

I criticize this excellent essay, and Dewey's other related writings, for two omissions. The first, of which I need say no more here, is that it implies a profundity in the understanding of elementary subject matter which teachers in fact are typically lacking, and in the development of which they need kinds of continuing education and practical support which our school systems — dominated in practice mostly by the first axiom, not the second — do not provide. The second and more basic criticism is that Dewey here, as elsewhere, neglects or fails to emphasize one central role of the teacher, one which when described will lead us to face the central topic of this paper, assessment in the service of teaching. It is a role which requires full acceptance of both axioms. Dewey was critical of many of his own followers (in the minority camp of progressive education) for not accepting the force of the first axiom, but he still reserved his big artillery for their opponents.

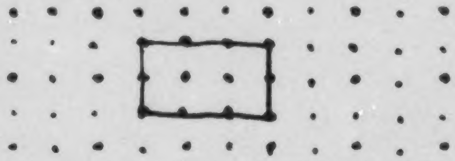
To put the point most sharply: In the essay referred to Dewey recognizes that a teacher's role is that of creating an ambiance in which "the child and the curriculum" are brought together in some fruitful matching relation, an ambiance which includes the teacher as an adult intermediary, as one who evolves that ambiance in step with children's development and learning, unpacking and reconstructing curricula in the process. Dewey has however nothing to say about the instructional role of the teacher _in_ such an ambiance, and so implicitly, in the end, gives support to those of libertarian or anarchistic persuasion who minimize that role in theory and neglect it in practice. How then should one conceive this instructional role, while having due regard for all the implications — as to the necessity of self-directed activity in model building — of the second axiom?

In that enriched ambiance which Dewey rightly conceives as a necessary condition for adequate education, children will have choices, and if the ambiance is well-designed and evolves well, these choices will be educationally significant ones. Recognition of the centrality of children's freedom to choose within such an ambiance is an easy consequence of the second axiom, and its advocates will often use the locution of "giving choices." The practical translation of this "giving," is often that in what nowadays would be called an open classroom there is a diversity of activities and materials available and "set out" for children to become engaged with, while a teacher is available, moving about to assist, to question, to encourage.

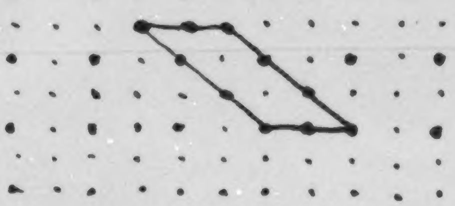
Desirable as all such provisioning may be, as a matter of course, one must question whether or how — though it is often desirable and too frequently lacking — it is really of the essence. Classrooms which appear on the surface to lack it may nevertheless be excellent, and those which provide it may fail. I believe the essence, from axiom II, is of a different order. Let me say: significant choices are _invented_ or _constructed_, they are very seldom simply "given." The process of choice is _part_ of the model-building activity, of learning, not something prior to it which can somehow just be "given" in the spirit of "here are the alternatives, you choose." Alternatives presented in this way represent superficial or conventional choices. At best they are _initial_ moves, moves designed to elicit information by a teacher, very seldom more than a potential doorway into subject matter or a source of steadying involvement and comprehension. In our own experience with early math and science we have seen many times that a rich array of enticing materials and phenomena will prove attractive to groups of children, in their own classrooms or in visits to our advisory center. On such opening occasions a laissez-faire attitude is for a time fully appropriate, one does not rush in to instruct — but it is not a long time typically, it is what one of us called "Christmas morning." If this is continued too long, one will begin to see the fading of interest, "running out of steam," the signs of boredom more often associated with conventional classrooms of too narrow a style. Cut off, on the other hand, by a "now let's get down to work" command, such an opening phase has little educative virtue, it is only a drop of nectar. The crucial phenomenon of significant choice comes rather from communication _around_ these early indicators of interest and readiness, an opening up of fresh alternatives for investigative curiosity, for the acquisition of skill, for the consulting of available sources; and above all, for extending initial involvements into those of longer term, in which what has been only a preliminary exploration is worked out, filed, and retrieved in later use.

A teachers' role in this process is that of helping children to find pathways of learning.(2) This role has two major aspects. One is that of assessment and planning, the other of investment, of joining, as an adult, in the pursuits being shaped and fostered, investing them, in the eyes of children, with adult enjoyment and dignity. I shall return to the first, as the main concern of this essay; but the second is important by way of background. The quality of a teacher's own perception of subject matter determines the frame or frames within which children's significant choices can come to definition, and is therefore crucial to choice. In part this range of potential choice depends on the teachers existing repertoire of available materials and their uses. If this is narrow and conventional, potential choices are limited as well. If it is wide, there is a greater probability that the teacher can help evolve fresh choices consonant with the beginnings which children will show themselves ready to make and extend. Since a teachers' subject matter range and understanding of subject matter limits that teachers' capacity for its investiture in clothing familiar and attractive, it also limits the teachers ability to assess and plan.

I give a small example. A second-grade teacher has been introduced to geoboards and has given these, with an ample supply of rubber bands, to each of 27 children for a lesson. They are first invited to play with the figures they can make. Then the word "rectangle" is discussed and illustrated, and finally everyone is invited to "make a rectangle" on this lattice of nails. The teachers' example has been a rectangle:



two high and three wide, and almost all now replicated it. We are going to "count the squares," and that will be our introduction to area. Each child is asked to count and most say "six." But one child has made a first rectangle this:



An observer saw him look about and — alas! — change his rubber band to the now-conventional form. But then again -- mirabile visu! — from some inner conviction and courage he changed it back again. When his turn came he therefore said, "six." The teacher, dutifully following instructions, not understanding, was disconcerned — something was wrong. Afterward the observer was able to pick up the neglected opportunity, and the way was opened for looking at this square on the diagonal and a more adequate approach to the concept of area. The child in question was ready for and delighted with this opportunity, and his work could have provided an entree to some real geometry for other children as well. But the lesson did not include any such opportunity.(3)

Given this sort of definition of a teacher's range — sadly inadequate in the case cited — one can then discuss matters of assessment and planning, within that range or extending it.

The situations presented by such opportunities is one which calls for a certain type of information matching. A channel is to be developed which gives access to subject matter for children who have given signals as to how that access may be achieved; among potential ways of access, some are suggested by those signals as promising. Starting from the other side, a match is to be achieved between some subject-matter topic or content and a diverse array of children with their available talents and resources.

However undertaken, the primary process of assessment is one I choose to tag, for specific reasons, by some behavioral, but emphatically non-behavioristic label. A relevant operative term is understanding. As teachers we wish to assess and chart for example, the progress of children's understanding of the unequal arm balance. The context I have in mind is work with some variety of materials such as weights, some identical with each other, some diverse; a long board to be balanced (or unbalanced); on a rounded support (for stability), sheets of hardboard to be balanced (or unbalanced) on a hemispherical surface, materials such as Tinkertoys to be assembled into arbitrary configurations to balance (or not), on a single point.

In the course of initial play with such materials children (and adults) will give behavioral evidence as to their understanding of what, for shorthand, we may call the law of moments and of stability or instability in equilibrium. Students' achievement of such understanding is our curricular objective. I shall say, however, that this objective is not to be exhaustively defined in terms of specific behaviors, as that term is usd in the recently (and still currently) fashionable notion of behavioral objectives. The latter notion is based on a philosphical or methodological opinion that the content of learning can be defined only in terms of objective data, some specific itemized list of specific verbal or performatory "behaviors," i.e. responses to such questions or commands as "place block A where it will balance block B." The listing may be long, but when set forth adequately it will give a behavioral or operational definition of the degree to which one has mastered "balance." Such a listing can then become, under Axiom I, a guide to the teacher, by which students can seriatim be taught not only general verbal responses but also correct performance.

In opposition to this view I put forward the view that understanding is the operative word; understanding is per se non-behavioral; on the other hand evidence regarding whether, or the degree to which, something like the concept of balance is understood, is behavior. The view rejected is a hangover from the logical positivism of the 1930's, and its verification theory of meaning. According to this epistemological theory a meaningful scientific statement is one which can be translated into the list of observable phenomena which can be said to verify it. The simplest refutation of this view — now almost unanimously rejected — is that the list of such directly observable phenomena corresponding to any hypothesis of scientific importance is always in principle inexhaustible. The hypothesis can and must be tested by observation, but is not defined by such evidence; if it were so formulated it would be useless, since all of its implications would then be exhausted.(4)

The operative meaning of "understand" puts this concept in the category of a term which cannot be exhaustively defined by any pre-determined list of behaviors. If we could train students to a criterion level of performance with respect to his understanding — so defined — it would not necessarily imply understanding, and indeed — if the training were sufficiently routine — might wholly miss the mark. Indeed — according to Axiom II -- our aim is that the student should build some model for the wide array of balance phenomena, one which is in some measure equivalent to the distillation of simple principles first enunciated by Archimedes; should not only build such a model but should be able to retrieve it from memory for use in diverse situations of a familiar kind, but also for trial in diverse situations, some of which are novel in aspect. The extent to which such models have been built, at any point in learning, and are retrieved in new situations, is testable in a teachers' observation, and it is from such evidence that the teacher can in turn attempt to build a model of the student's model by comparison with the teachers' own model of phenomena — in this case balance.

Well-constructed models have a characteristic power (5)to reduce the redundancy of experience. Behavior can exemplify the use of a model and give clues to its nature, but in its own nature a model is of a different order. In its own nature a model is first a way or habit of selecting, organizing and providing information, and then later -- by abstraction — an object in its own right, a conceptual reality which we can describe and analyze — e.g. the law of moments, the conditions of stability — in the language of physics or mathematics, not the language of human behavior per se, though as a retrievable model it must be richly indexed to phenomenal and behavioral imagery.

Understanding, so conceived, is, in principle, never complete. Models in this sense can become linked to other models in a network, thus further reducing residual redundancy 2.3*, and increasing what might be called the cross-section area of possible applications ("transfer of learning"). So the conceptual frame of balance may be linked to that of mechanical work and potential energy, or to other cases of the use of an efficacious center (Holton), to geometry (Archimedes), and so on. In another direction it might become linked to the barometer and the ocean of air, to still other phenomena of equilibrium and to the image of the potential well, etc.

The representation of understanding by the idea of a growing network serves also to suggest why there is wide latitude for educational choice in the time-ordering of many specific topics, at least at early levels of learning. Important ideas — frames and modes of understanding — are met along many tracks of learning, that is why they are important, and that is why subject matter is open to reconstruction for learning in many ways.

From this assertion of the adaptability of subject matter I turn to the other pole, the adaptability of children. It is only when these two kinds of adaptability are seen in conjunction, I propose, that the child and the curriculum can be fully brought to harmonious relation.

To begin the discussion I propose to introduce two subsidiary lemmas about the assessment of ability. One that if we are to speak about measures of ability or talent, in the biographies of any individual at any time, this measure should be conceived as a vector of many dimensions, not just one aggregate (e.g. I.Q.) or a few (e.g. the subsections of the individual I.Q. tests). The individuality of learners implies it, it is a theorem about the career of the human model builder.

It is practically confirmed by the fact that in any group the rates of learning along any one curricular track are conspicuously different, and there are conspicuous changes (often inversions) in these rates as a function of the kinds of ambiance, access, and teaching involved. The second lemma is that learning rates are roughly proportional to relevant antecedent leaning. The first lemma implies a profile, a vector of abilities and talents (which I visualize in polar coordinates) of which no single function (average, etc.) is either very meaningful or very useful in teaching. The second implies that in any given specific direction on the polar profile the distribution of abilities in a group should be something like the log-normal distribution, with a large variance between individuals.

Under these lemmas it will follow, most importantly, that the assessment of strengths — peaks of background, skill, knowledge, talent — is of prior importance to what is also necessary, that of weakness, low points on the profile. Thus a child with visual-artistic strengths has a different potential for access to geometry or arithmetic or reading from one with special verbal or mechanical facility. Since rates of learning are dependent on learning already achieved, the potential for bridging over from an existing strength or talent to overcome weakness, alone. But here the role of the teacher is paramount, in finding ways of building cross-over linkages between areas of strength and of weakness, and thus helping children to find choices which are both attractive and educationally significant.

These two lemmas, I believe, indicate the principal reasons why prevailing ideas of formal assessment are of very limited use in teaching, and often are damaging. As to the positive side. Scores on such tests are typically a confirmation of what teachers do or should already abundantly know. A child who has become seriously addicted to reading will before long far exceed, in score, the age norm for reading ability derived from standardization of such tests. The same is true of arithmetic. To demonstrate reading levels slightly above these norms may comfort a teacher, but it is surely no sign of excellence in children's work. Moreover to aim instruction at the typical content of such tests is in most cases to substitute routine skill training for the more basic art, that of investing reading with value for children in relation to their expanding interests in the world around them, in fantasy and story telling, in writing of that which they deem worthy to tell of their own lives and learnings.

If the above outline of the desiderata of successful teaching is accepted, then one has a background for the selection or invention of specific means of assessment which such teaching requires. A first consideration is that time scales, the characteristic return-time from assessment to its uses in teaching. These vary from minutes to months. Records (in memory or on paper) are vital because the way assessments influence teaching needs to be monitored by the teacher; individual decisions in teaching are fallible, and their success or failure should confirm or modify teachers profile models of individual children and should contribute to the teachers' own professional growth. The design of professionally useful techniques for assessment and record-keeping must come however as a harvest from successful practice and is unlikely to be provided by professional test-designers unfamiliar with the needs of the teaching art. I suggest that we should examine examples of such techniques when we can find them proposed or in use.

Given what I have said above about the multivariate and log-normal distribution nature of such data, they are unlikely to resemble formal test scores, though they may sometimes incorporate such measures. It should be remembered in this connection that any reliable yes-no discrimination _is_ a measurement, and that where the number of dimensions of interest exceed the number of data such discriminations are likely to take the form of a paragraph than of a number.

As to assessments of a more long-term relevance in teaching, my theorems and lemmas do not exclude formal tests — standardized or not — as sources of confirmatory evidence useful to teachers. If my argument is correct these by themselves — though of limited usefulness — _can_ be usd to sample children's learning and skill in subject matter areas, provided they do not get confused with more significant ways of defining the aim of education. They can sometimes reasonably be considered as necessary conditions of educational success, but they by no means should be confused with what is sufficient or — to use a currently fashionable term — basic.

NOTES

1.John Dewey, "The Child and The Curriculum."

2. In what follows, I am especially indebted to Frances Hawkins, though she should not be held responsible for my generalizing interpretations. cf. THE LOGIC OF ACTION, Pantheon, 1974, and "The Eye of the Beholder," in SPECIAL EDUCATION AND DEVELOPMENT, S. Meisels, ed., Univ. Park Press, Baltimore, 1979.

3. See also Frances Hawkins, "The Eye of The Beholder"

4. cf R. B. Braithwaite, Scientific Explanation, Oxford, 1963

5. for a formal definition of "power" in this sense see D. Hawkins, "On Chance and Choice," REVIEWS OF MODERN PHYSICS, vol. 36, no. 2, April, 1964, pp. 512-517.

6. cf. D. Hawkins, "On Understanding the Understanding of Children," in D.H. The Informed Vision, Agathon Press, 1972, and at another level, Edwina Michener, "Understanding Understanding Mathematics,"

INTRODUCTION TO PART III

In the discussions of our panel several themes emerged time and time again with great forcefulness. The issues these themes dealt with were of two sorts.

The first kind of issue raised was that of the constraints that present institutional structures and organization place on possible alternative assessment practice. The second kind of issue raised was the nature of the desirable features and properties of new alternative assessment practice. The three papers that follow, by Parker Damon, Asa Hilliard and Howard Gruber & Robert Keegan address these issues directly.

J. Parker Damon is principal of the McCarthy-Towne School in Acton, Mass. He writes from the perspective of a practicing school principal. That perspective is augmented and complemented by his experience as a Ford noundation Fellow with project TORQUE at the Education Development Center during the 1977-78 school year, and his participation in the 1979 National Institute of Education conference on Testing, Learning and Teaching.

Dr. Damon believes that schools and teachers have all too little of a precious commodity called time. Thoughtful instruction and sensitive assessment take time. In the first part of his paper he shows how the time demands of present assessment practices cut deeply into teachers' available time, without the compensation of yielding useful information in return.

In the second part of his paper, he outlines some assessment practices that are both alternative to, and complementary to standardized testing. In this section, he draws heavily on the ongoing experience of his own school as well as the experiences of other educators with whom he is in close and continuing contact.

In the last part of his paper, Dr. Damon discusses the several sorts of support necessary to change practice. In particular, he points out that not all problems are solved by throwing money at them. Some sources of support are there for us to use without further expenditure of funds. These new sources of support involve the introducing of new actors into the educational scene in the form of parents and older students. They involve the encouragement of teachers' professional activities and development. Above all, they call for a more realistic and informed view of the realities of schools and teaching.

Asa G. Hilliard is Dean of the School of Education of San Francisco State University. He writes from the dispassionate perspective of the scholar and from the impassioned perspective of one deeply committed to social change in the United States. This counterpoint of perspectives recurs continually throughout his contribution to this volume.

The thread that ties Dean Hilliard's paper together is the celebration of diversity. People differ from one another as individuals. When they form into groups, either under their own volition or under pressure from others, the groups they form differ from one another. Jerrold Zacharias once said, "children are different from one another, and schools should make them more so." Asa Hilliard clearly subscribes to this view.

Dean Hilliard begins his paper by an inquiry into the nature of culture and explores some of the reasons that current assesment practice is as insensitive as it is to cultural variation and diversity. He goes on to examine the meaning of the term "test" in education and the interacting triad of considerations of the type of test, the use of the test and the user of the test result. All too often, schools and society have paid dearly for the confusion of these considerations in the minds of the public. Finally, in closing his paper, Dean Hilliard draws up a list of guidelines for the shaping of new assessment practices and instruments are very much in the spirit of the other contributions to this volume.

In the long run one of the goals of education is to have students internalize the assessment function and reflect on the quality of the own learning and doing. Indeed, leading an "inspected life" may well be regarded as the hallmark of a successful education.

By and large we don't devote much effort in our formal educational systems to helping students develop the ability as well as the inclination to do this. Howard Gruber and Robert Keegan, of the Institute for Cognitive Studies at Rutgers University, describe a course in psychology they offer to non-traditional students that emphasizes the importance of reflection on ones own thought and learning and offer some explicit suggestions drawn from their experience to help those that seek to move in this direction.

INVESTIGATIVE TEACHING: AID TO THE STUDENT

by

J. PARKER DAMON

McCarthy-Towne School
Acton, Mass.

WHAT PREVENTS TEACHERS FROM MAKING EFFECTIVE USE OF STANDARDIZED TESTS, OR OF DEVISING AND USING ALTERNATIVES TO THEM?

Standardized tests have an impact on curriculum content, budget priorities, and faculty assignments. They are used to identify individual students for inclusion or exclusion in special programs. They influence teacher behavior. Sometimes this influence is great, sometimes not. Whole units of study may be added to or deleted from the curriculum; time allotments devoted to a particular activity may be altered; sequences of learning experiences may be switched. As result of poor performance on a language mechanics subsection of a test, a district or school may purchase a whole new series of language arts textbooks. Teachers may be told to spend more time on this area of instruction in isolation as opposed to integrating the teaching of grammar, punctuation, usage and capitalization with the students' other work on reading and writing. A weak showing on the study skills subtest may pressure a teacher to revamp the curriculum so that students will have to use resource books in place of other reseach activities. Every one of these influences work in the direction of further constraining the time the teacher has available.

The amount of instructional time available to teachers for whole class projects, large group instruction, and sequential small group activities, is not as extensive as some may think. After daily organizational meetings, lunch, recess, physical education, art, music, and special classes for certain students are deducted from the twenty-seven and a half hour school week, not much is left. For example, during a typical week the time not available for whole class instruction (i.e., all students present in the classroom at the same time) might include: 1/2 hr/day for morning meeting and predismissal cleanup = 2 1/2 hrs/wk; 1 hr/day for lunch cleanup, lunch, lunch recess = 5hrs/wk; 1 hr/week for art, music, physical education = 3 hrs/wk; 1 hr/day when some students are out of the room for special classes = 5 hrs/wk; 1/2 hr/day for recess or other kinds of recreational activity = 2 1/2 hrs/wk; 1 hr/week for unexpected miscellaneous activities. The teacher may have eight and a half hours per week when all the students are present. These hours, however, may not be available in coherent blocks of time or at the most advantageous times of the day or week. Thus when a teacher is faced with making the best use of both the nineteen hours when not all students are present and the eight and a half when they are, it is not surprising to find other pressures or incursions having a marked impact on teacher attitudes and behavior.

Integrating the different curriculum areas, planning developmental activities that have a long range purpose, providing for exploration, experimentation, discussion, are components of the instructional process that require a lot of teacher effort and a lot of available instructional time. Pressures from the outside in the form of standardized test outcomes will, whatever their merits, force other priorities, activities, materials or methods to give way. In this way tests can have a direct impact on classroom instruction that the teacher may, or may not, agree with.

In addition, there may also be indirect kinds of impact that are not appreciated at first. Often someone or group other than the classroom teacher believes the test results signal something different is required. Administrators' assumptions, parents' perceptions, and citizens' concerns may pressure the teachers to do what the teachers know to be unnecessary or wrong, or warranted but poorly timed, or appropriate and doomed to fail (because the requisite support is lacking). Impacts of these sorts are second-hand, indirect, and delayed.

The daily instructional process is, in the main, unaffected by the information produced by standardized tests. To the extent that there is an impact, it is usually a negative one. Eva Baker points out that "studies show that what teachers do as a result of test scores is to drop whatever it is they are working on and do something else, or to repeat what they are doing more frequently. Neither of these are examples of positive or constructive use of test information. Teachers are not using the information to "open up their instructional repetoire."

Teachers have a variety of good reasons for not using the data produced by standardized tests. However, these reasons are often overlooked by those whose views of teachers are based on faulty stereotypes, emotionalism, and myth. It is, therfore, important to point out that;

> —Teachers are not lazy. Although the school day and year are short, their work time more than equals that of others who work 9 to 5 for 50 weeks of the year.
>
> --Teachers want to be accountable for their performance. But they also want and deserve support so that accountability is an avenue for professional improvement rather than simply an avenue for blame.
>
> --Teachers welcome assistance intended to improve students' specific learning experiences. Successful materials and practices are always being sought. Thus anything which is easy and effective in terms of providing teachers with accurate, insightful, diagnostic, relevant, immediate, concrete, complete, and constructive information would be welcomed. Critics of standardized tests argue that none of these criteria is met by any group administered achievement test.
>
> --Teachers are willing to devote extra time (however defined) to improve the learning experiences of their students. This willingness includes becoming more proficient in the use of tests and other assessment practices. The fact is, however, following participation in workshops and courses designed for this purpose, the active, ongoing positive use of standardized tests does not increase.

Why then, do teachers use these instruments so infrequently?

There are several reasons why teachers fail to use test data. First, there are the many mechanical impediments that get in the way. Turn-around time from when the student takes the test to when the teacher receives the results is too long. To be useful for instruction, that time ought to be no more than a day or so. Usually, several weeks to a month pass and frequently there are unexpected delays. Test items bear marginal resemblance to daily classroom work. The relationship of test goals to teacher goals and to each teacher's sequence of instruction to reach them occurs only by chance. Moreover, the information provided to the teacher is usually too sparse or too superficial or both for it to be of use even if it arrived promptly and related to what the teacher was teaching.

A second reason for the non-use of test results comes from the constraints of the materials schools and teachers must contend with. In determining what instructional materials they and their students will use, schools and teachers usually have only two choices. Teachers may buy them from suppliers whose wares are practically indistinguishable, or make them themselves at night, on weekends, or during vacations. The latter path is demanding. Adapting, collecting, creating are time consuming efforts. It is unrealistic to expect teachers to discard what they have created and believe to be worthwhile on the basis of information which they do not value much.

The argument that teachers will make better use of criterion referenced tests (than they now do of norm referenced tests) because they can participate in selecting test items fails in face of the fact that these items are usually selected to represent a district's goals and not to reflect what students are doing and learning in the classroom. Teachers use materials in idiosyncratic fashions that usually make standardized test information inappropriate for assessing student performance in the classroom. Some instructional management systems attempt to get around this difficulty by means of intricate crossreference and index schemes, and detailed sequence charts and goals checklists.

A third reason teachers do not use standardized test results is that they have too many students to work with WHEN TEACHING RESPONSIBILITIES ARE CONSIDERED ALONG WITH THEIR OTHER RESPONSIBILITIES. Even if a teacher knows how to interpret test results and how to translate them into learning experiences, it is unlikely the teacher will have the time to do so for every child on a subtest by subtest basis. Even less likely is an examination and comparison by the teacher of the individal test items and responses for each child. As a result, the teacher must rely on the summary printouts showing which items were correctly and incorrectly responded to, the frequency of errors of individual students compared to their classmates, scores of one kind or other compared to what might be expected (anticipated scores) and to the scores of norm groups. Scanning numbers on computer printout sheets is quicker than looking at each individual student's answer sheet and comparing responses to the actual questions which were asked. It is also more superficial and further removed from classroom activity and direct intervention in the teaching-learning process. Even if the teacher is sophisticated and knowledgeable about how to use the information provided by summary printout sheets, other time pressures such as preparing daily lesson plans in four or five curriculum areas, working on curriculum development projects for school or district, responding to parent and community concerns, and working with specialists in order to attend to students with special needs may well take precedence. Those who work in schools, like most everyone else, do not always have the luxury of adequate time for all that has to be done; shortcuts are used even if they do not improve the quality of what is being done. If the teacher is in any way unsure of what the test results may mean, or how they may be translated into classroom practice, there is little likelihood of their being used.

ILLUSTRATION 1

METHODS FOR EVALUATION PROGRAMS AT McCARTHY-TOWNE

1. Parent Information Coffees
    a) At school
    b) In neighborhoods
2. Parent-faculty annual meeting
3. Parent-faculty-student surveys and questionnaires
4. System, state, national tests and surveys
5. Reports by graduate students
6. Faculty and school self-evaluation
7. Interviews of Sixth, Seven, and Eighth Graders
8. Surveys of Junior High School faculty
9. Creation of school's own subject area objectives
10. Creation of school's own tests of objectives
11. Attitude scales and inventories
12. Survey parents of children who once attended McCarthy-Towne
13. Collection of comments and concerns from public
14. Observations of student teachers
15. Reactions of results of evaluation data from all concerned
16. Videotapes of school's programs in action
17. Samples of students' work
18. Third-party evaluators
19. Comments of vistors
20. Reactions from other schools and professionals

WHY DON'T TEACHERS DEVISE AND USE ALTERNATIVES TO STANDARDIZED TESTS?

Teachers do create their own assessment instruments and procedures. Teacher made worksheets, samples of student work, professionals' anecdotal logs, end-of-the-chapter or unit tests in textbooks, and many other forms of assessment (see Illustration 1) exist and may be found in poor and affluent schools in urban, suburban, and rural districts. In many school systems though, such alternatives are distrusted. As a result, schools and districts operate on a two track assessment system. Assessments intended to assist administrators and school boards in making policy, priority and program decisions depend on standardized tests, while assessments made to improve classroom instruction and individual student performance depend on a variety of techniques.

> Some people might call such information more subjective, and thus more suspect, than test scores. Those who make this charge should be reminded of the highly subjective nature of test construction, to say nothing of the interpretation and use of test data. The issue may not be whether to use these alternatives, but whether the person requesting the information trusts the one providing it. Right now, the level of trust and confidence between the public and the professionals throughout the country seems low.
>
> (Damon, J. Parker, "Questions You Should Ask about Your Testing Program," The National Elementary PRINCIPAL, Vol. 56, No. 1, September October 1976 p. 53; also reprinted in THE MYTH OF MEASURABILITY, Paul L. Houts, ed.)

Teachers trust the results of other assessment instruments and practices more than they do the standardized tests. In addition to the uneven quality of information provided by standardized tests, they also tend to continue other confusions. For example, they encourage the use of labels or terms such as "measurement," "assessment," evaluation," standard of performance," literacy skills," "basic skills," "hierarchies of skills, abilities, and thinking," "knowledge," "understandng," "attitudes," "aptitude," "anticipated achievement," "grade level performance," and many more as if they each have precise meanings that distinguish one from the other or apply with certainty to both groups and individuals. In reality, and more often than not, these terms and labels disguise ignorance and promote myth. (See Illustration 2)

ILLUSTRATION 2

Investigative Teaching Model

Students, parents, teachers, administrators, and members of the public all want to know how well performance, task and situation, and goal match. But each may want this information in the perspective of one looking back into time to examine completed performance, from the vantage point of one observing ongoing activity, or as a predictor of future achievement. Why did someone do that? Why is someone doing this (rather than that)? Why will someone be able to do that? When we look into the past, we are evaluating. Observing present activity is assessing. In looking to the future, we are estimating the likelihood of a prediction being realized. Evaluation, assessment and estimation are words to use carefully, not interchangeably, for they have different meanings.

PAST
EVALUATION
QUESTIONS

What did the person do?

Was the goal reached?

How well was the goal reached?

Should the outcome have been different?

How could improvements be made?

Why should the activity stay the same, be changed?

PRESENT
ASSESSMENT
QUESTIONS

Why is the activity being performed?

Is it being done the way it should be?

Are the appropriate people, materials, and conditions involved?

Is intervention appropriate?

Where is the activity headed? How many outcomes?

FUTURE
ESTIMATION
QUESTIONS

Will the person be able to do the activity?

How will we or the person know before, while, and after doing it?

How can we help prepare the person for doing the activity? Should we?

Why will one form of assistance be better than another?

The classroom teacher evaluates, assesses, and estimates all the time. Standardized tests, chapter tests, and other commercially available instruments provide only some helpful data the teacher may use to answer the questions in the first column. Teachers have to rely on their instincts, their powers of observation, their tailor-made or chosen materials, and their skills of directing and demonstrating in order to promote the learning situation. The teacher has to investigate what is going on firsthand to be able to answer the questions in columns two and three.

WHAT WILL IT TAKE TO HAVE THE PUBLIC, POLICY MAKERS, AND OTHER PROFESSIONALS TRUST THE JUDGEMENT OF TEACHERS MORE, AND THE RESULTS OF STANDARDIZED TESTS LESS?

Attacking the credibility of standardized test results or the tests themselves will not cause a change in faith. Probably no single course of action will inspire greater confidence in teachers' judgment. However, it is likely that a series of concerted efforts would have this result. First of all, teachers need to know what they are talking about when discussing student performance with others. They have to be informed about the strengths and weaknesses of different assessment instruments and procedures, about how their classroom's curriculum content, their students' learning styles, and their own assessment practices, and about how their classroom work supports the overall objectives of the school. If teachers are able to articulate these relationships clearly, and if they hold themselves and their colleagues to agreed upon standards of performance related to integrating assessment, instruction, and school goals, then teachers are more likely to trust themselves and be trusted by others.

Teachers require inservice training to reach this level of understanding and confidence. Few teacher training institutions instruct teachers-to-be on how to evaluate the content and appropriateness of assessment and teaching materials. (1) Summer workshops and released time during the school year for follow-up seminars and information exchange sessions could provide teachers with the knowledge needed to be comfortable and confident when communicating their judgments, and to implement their judgements constructively.

Supervisors, principals, and curriculum specialists who have the responsibility for interpreting pupil performance to the public need inservice training as well in order to help them put assessment information into the proper perspective by emphasizing how such information should be interpreted.

WHAT SUPPORT IS NECESSARY IF TEACHERS ARE TO ENGAGE IN USEFUL ASSESSMENT PRACTICES TO IMPROVE INSTRUCTION?

There are different kinds of support, different times support is needed, and many variations for combining the timing and type of support. Talking about providing teachers with support of various kinds is easy; providing it is something else. Support comes in the form of money, the time to do things, encouragement and reinforcement from colleagues and supervisors, the flexibility to change schedules and activities, the space in which to plan, operate, and store, the services of curriculum consultants, classroom assistants, and clerical aides, the opportunities for continuing inservice training, and access to many different means of communication. These are the types of support a district or school can give its faculty. Few, however, provide more than limited amounts of any one of these supports. Fewer still provide any of them for any length of time. There are just too many jobs to be done, too little time in which to do them, and too few resources. In one unusual instance, a district not far from Boston made the commitment to a long term multifaceted effort to improve instruction via the continuous use of assessment. (See the NESDEC booklet describing the 10 year Fitchburg project.) But most districts or schools are unwilling or believe themselves unable to alter priorities in order to provide the support teachers need to improve instruction on more than a hit-or-miss basis.

Beyond the schools' support of teachers is that of society. Foundation endowments, government grants, regionalized and collaborative local efforts make the establishment of teacher resource centers, inservice institutes, materials and resources exchanges, experimental and dissemination sites, and information networks possible and practical. The boulder representing the support for teachers to improve their instruction through the better use of assessment is poised, ready to be rolled down the hill of practice. Teachers want the support, administrators want to provide it, and the public is beginning to recognize that much as it is in the other professions, improved tools are necessary but not sufficient for long term improvement and reform.

Many people believe that of all the kinds of support teachers require - money, time, encouragement, autonomy, flexibility, space, people, inservice, and communication -money is the most important. I am not so sure. I think that perhaps encouragement is the key element to a successful support system. Encouragement can be in the form of another professional describing and defending what colleagues are doing. Encouragement can be the recognition of the importance of a job to be done, the commitment to it and the work of others to get it done, and the development of a similar recognition and commitment in others. Refocusing curricular emphasis, changing curricular content, improving instructional practices, and restructuring learning experiences are all worthwhile efforts most schools are concerned with pursuing. But they cannot all be done simultaneously, and well. Which comes first, and how to support this developmental activity requires a long term commitment to carefully established priorities. In this sense, encouragement and commitment are synonyms.

Money is, of course, an obvious and necessary form of support that makes other forms easier to have. If teachers are to use assessment to improve instruction, then they are going to need materials for use with students before, during, and after the assessments are made. It is quite likely that many of their existing materials will have to be modified or supplemented. Teachers and other faculty members will also need time to learn about alternative assessment practices and tools and about how to link assessment to the improvement of instruction.

Time will be required during vacations, at the end of the school day, and as a result of being released from regular responsibilities. The more frequent use of substitutes or the provision of other forms of "classroom coverage" (via volunteers, older students, placement in other classes, or alternative educational experiences e.g., parents or neighbors supervise students' experiences away from school (creative hooky) are a necessary form of support so that teachers may attend inservice workshops and planning sessions.

Released time during the school day is necessary if teachers are to use assessment to improve instruction in a serious way. The time I am referring to should not be confused with the planning or preparation periods many teachers have. These periods typically occur when students leave the classroom for art, music, or physical education. Though they may be used to reorganize the next series of activities on the basis of what has just happened in the classroom, they are more likely used to organize materials, correct papers, or catch up on communications with colleagues and parents. Other faculty are usually not free at the same time so joint review and planning is not possible. Periodically, teachers need additional time in order to contemplate the information provided by assessment. Such time should be found during the school week. To relegate the review and use of assessment information to after school, weekends, and school vacation periods is to mistakenly believe these instructional improvement activities can be put off and still be useful.

---

(1) Fred M. Hechinger, "About Education, Study Suggests Texts Are Often Inadequate," The New York TIMES, April 8, 1980, C4.

CULTURAL VARIATION AND LIVING ASSESSMENT IN THE SERVICE OF INSTRUCTION

Asa G. Hilliard III

> "Man has put himself in his own zoo. He
> has so simplified his life and stereotyped
> his responses that he might as well be in
> a cage."
>
> (Hall, 1977)

Historically, standardized testing as it has been used in education has reflected users' strong committment to sorting children, guessing at or predicting children's future performance, measuring "school achievement", and "diagnosing" learning difficulties. Further, to accomplish this there has been an unwritten but strong demand for mass produced universal instruments which could be easily administered, quickly scored, and inexpensive. It is this peculiar combination of things which has impeded educators and researchers in the search for tests or assessment procedures which can be shown to make a positive difference for learners in the educational process. It is a pity, since testing and assessment can be systematic, rigorous and, above all, valid without being standard and universal. But most important, testing and assessment, appropriately constructed and conducted, can and should make a positive difference for children in their education.

All people "swim" in culture. Culture is the stuff that people make. At the basic or "deep structural" level, people all over the world appear to perform similar functions. They construct language and learn language. They organize and classify their experience according to the ways that they have created. They expand their repetoires to accomodate and assimilate new experiences. They do many other "cultured" or people-made things, but they don't all look the same or do things in precisely the same way. At the surface structural level, they manifest their common equivalent human basic competencies in a variety of ways.

A few years ago, the loss of culture phobia and academic recognition of cultural variation in the testing area led to attempts to imagine how people would behave if culture were held constant. This resulted in a "culture free" testing movement. As it has become more and more apparent that the very question that an examiner asks is itself a bit of culture, not nature, the goals of testing have begun to reflect the idea of "culture fair" testing. However, neither "culture free" nor "culture fair" testing as we now know them seems to have much academic meaning or practical utility. The problem for educators is neither to do without that which all people must display (culture) nor to test by providing an equal number of items for each culture or items which do not favor one culture over the other in the final score (culture fair). Rather the problem for educators is to use culture boldly as the medium of communication and creativity.

It is my intent to illustrate the value of professional practice of a skillful use of culture - the stuff within which we do indeed swim, without which being human would be meaningless if not impossible, and in ignorance of which pedagogy is a joke. Specifically, I intend to treat several key issues:

1. How can a knowledge of culture prevent diagnostic error?

2. How can a use of culture reveal material for more effective planning of valid instructional strategies?

3. How can a knowledge of specific cultures and an understanding of the concept of cultural variation serve as a basis for constructing tests which do not confuse quality achievement with cultural myopia?

4. How can the use of culturally specific tests and assessment procedures assist teachers to help children to construct expanding repetoires?

5. How can the use of culture enable children to enter into dialogue with their mentors and to assume their responsibilities as learners - as culture creators?

6. How does culturally sensitive testing and assessment allow for a more valid approach to "accountability", or, put another way, help educators and others to know what has happened in the learning process and how it happened?

Sophisticated testing and assessment in the service of instruction which uses culture is already in operation. Consistent and dramatic learning is seen with learners who, by traditional mass produced tests, would be classified erroneously as unable to learn much or as having learned too little to make the next step in teaching worthwhile. It may not be mass produced, universal, or cheap, but it can be valid and useful for improving instruction.

THE REALITY AND MEANING OF CULTURE

Typically, analyses of tests for cultural bias are accomplished by comparing the differences in the pattern of responses of two or more presumably different cultural groups to a common set of test items. Such an approach can shed little light on a very complex matter, primarily because it takes for granted that culture has been defined scientifically. It also allows the attribution of a cultural identity to subjects by a given researcher. The meaning of culture and the placement in a cultural group are critical matters for cross-cultural researchers. These matters are far from easy under any conditions. This is especially true within the United States of America, since cultural patterns may be either relatively distinct or they may be amalgamated or overlapping among groups. In any event, one cannot assume that culturally specific data be handled adequately by those who have not studied culture systematically and professionally. Without such background, there is a strong likelihood that the wrong questions will be posed and confounded answers obtained.

Cultural bias in testing will produce inequity for some groups because of error in assessment. But it is also just as important to think of culture, not as an impediment or threat to evaluation, but, as a prime ally in the testing and assessment process. Cultural experiences are data which can be used in testing and assessment. Indeed, they must be.

But what do we mean by culture? Although there is much variation among specialists in definitions of culture, there are common themes which run through the definitions. These themes come from systematic empirical observations of human behavior in natural settings. (Cole and Scribner, 1974) (Hall, 1977) (Levi Strauss, 1966) (Labov, 1970) (Ben Sidran, 1971). For example, Edward Hall (1977) illustrates something that he calls "extension transference" (ET). It is here that investigators impose their order on the reality of other people.

> "Another frequently dysfunctional characteristic of ET systems is that they can be moved around and inappropriately applied. This is understandable, because it takes years and even lifetimes to develop a good extension system. (Sometimes we call them paradigms when they take a grammatical or rule-making or modeling form.) In the days following the opening of Japan to the outside world, American missionaries wrote their own grammars for teaching Japanese to each other. Anyone who has seen one of these early grammars knows that the missionaries projected their own, Indo-European grammatical forms onto Japanese without any reference to the actual structure of the Japanese language. Nominative, genitive, dative, and ablative cases all appear in te grammars with identical Japanese words under each. A characteristic of transference phenomena is that people will treat the transferred system as the only reality and apply it indiscriminately to new situations. I once knew an American woman in Tokyo who became so resentful of the Foreign Service Institute language drill designed to reinforce the learning of proper Japanese that she simply struck out on her own. She said, "The devil with all these honorifics. I'm not going to learn them. I will simply learn vocabulary." What she spoke, of course, was a most dreadful, unintelligible melange of Japanese words and English grammar.
>
> Something similar has happened to significant blocks of social science. Not only has there been extension transference (not data, but methodology is thought of as the real science), but because physical science has been so successful, the paradigms of the so-called hard sciences were transferred intact to social science, where they are seldom, if ever, appropriate."

p.33

The inability of the investigators to understand that his or her own logic is not unique has impeded scientific discovery for many years and in many places. Claude Levi-Straus (1966) gives us many excellent examples of culturally specific logic.

> "Following Griaule, Dieterlen and Zahan have established the extensiveness and the systematic nature of native classification in the Sudan. The Dogon divide plants into twenty-two main families, some of which are further divided into eleven sub-groups. The twenty-two families, one of which is composed of the families of odd numbers and the other of those of even ones. In the former, which symbolizes single births, the plants called male and female are associated with the rainy and the dry seasons respectively. In the latter, which symbolizes twin births, there is the same relation but in reverse. Each family is also allocated to one of three categories: tree, bush, grass; Finally, each family corresponds to a part of the body, a technique, a social class and an institution (Dieterlen I, 2).
>
> Facts of this kind caused surprise when they were first brought back from

Africa.

(Levi-Straus, p. 39.)

Levi-Straus gives another example from American Indian Culture.

> The Hopi, like the Zuni who particularly engaged Durkheim's and Mauss's attention, classify living creatures and natural phenomena by means of a vast system of correspondences. The facing table is based on the information scattered in several authors. It is undoubtedly only a modest fragment of an entire system, many of whose elements are missing.
>
> p. 41.

THE LOGIC OF TOTEMIC CLASSIFICATIONS

	NORTHWEST	SOUTHWEST	SOUTHEAST	NORTHEAST	ZENITH	NADIR
COLORS	yellow	blue, green	red	white	black	multicolored
ANIMALS	puma	bear	wildcat	wolf	vulture	snake
BIRDS	oriole	bluebird	parrot	magpie	swallow	warbler
TREES	Douglas fir	white pine	red willow	aspen		
BUSHES	green rabbit brush	sage brush	cliff rose	grey rabbit brush		
FLOWERS	mariposa lily	larkspur	castilleja	anogra		
CORN	yellow	blue	red	white	purple	sweet
BEANS	French bean	butter bean	dwarf bean	lima bean	various	

> ...These are only a few of the examples which might be given. There would be even more examples than there are, had ethnologists not often been prevented from trying to find out about the complex and consistent conscious systems of societies they were studying by the assumptions they made about the simpleness and coarseness of 'primitives'. It did not occur to them that there could be such systems in societies of so low an economic and technical level since they made the unwarranted assumption that their intellectual level must be equally low. And it is only just beginning to be realized that the older accounts which we owe to the insight of such rare inquirers as Cushing do not describe exceptional cases but rather forms of science and thought which are extremely widespread in so-called primitive societies. We must therefore alter our traditional picture of this primitiveness. The 'savage' has certainly never borne any resemblance either to that creature barely emerged from an animal condition and still a prey to his needs and instincts who has so often been imagined nor to that

> consciousness governed by emotions and lost in a maze of confusion...
>
> p. 42.

A final example from Levi-Strauss should clinch the point.

> When he began his study of the classification of colours among the Hanunoo of the Philippines, Conklin was at first baffled by the apparent confusions and inconsistencies. These, however, disappeared when informants were asked to relate and contrast specimens instead of being asked to define isolated ones. There was a coherent system but this could not be understood in terms of our own system which is founded on two axes: that of brightness (value) and that of intensity (chroma). All the obscurities disappeared when it became clear that the Hanunoo system also has two axes but different ones. They distinguish colours into relatively light and relatively dark and into those usual in fresh or succulent plants and those usual in dry or desiccated plants. Thus the natives treat the shiny, brown colour of newly cut bamboo as relatively green while we should regard it as nearer red if we had to classify it in terms of the simple opposition of red and green which is found in Hanunoo.
>
> p. 55.

Thus we see that categories and classificatory schemes are not nature but culture. It is the "invisibility" of ones own culture which clouds the perception of observers, which impedes scientific progress, and which contributes to diagnostic error.

But let us return to a more articulated definition of culture, a definition which makes culture amenable to empirical investigation.

Every person or group of people is born into a unique environment. A part of that environment has been created through forces which operate without the conscious effort of people. Another part of every human environment is there as a consequence of the creative acts of people. It is this latter part, human creativities, that can be referred to as culture. To be even more precise, we may think of the range of creativities as including such things as the following:

1. Making tools such as:
   a. language
   b. levers
   c. categories
   d. symbols

2. Making esthetic experiences such as:
   a. music
   b. poetry
   c. art
   d. humor

3. Making history such as:
   a. stories
   b. documents or records

4. Making explanations such as:
   a. philosophies

   b. religions
   c. theories

5. Making values such as:
   a. mores
   b. ethical principles

6. Making rituals such as:
   a. holidays
   b. celebrations
   c. ceremonies

7. Making futures such as:
   a. expectations
   b. forecasts
   c. designs

8. Making government such as:
   a. order of authority
   b. laws or rules for conduct

In short, it is the unique patterns or configurations of all of these things which cause a group to be seen as sharing a culture. Advertisers know this and are able to target their sales appeal to particular cultural audiences.

> For example, a cardinal rule of postioning begins with the rank of products or brands on the ladder in the consumer's mind. It is foolhardy to advertise head-on against the No. 1 product or brand, because your advertising tends to reinforce the leader. This fact of life is even more significant among blacks, because they are more rank-conscious and they use rank for more deep-seated reasons. More than whites, they tend to select brands in the No. 1 positions and to use them as signals to their peers--and to whites...
> They are not good prospects for boats. Nor would they find much identification with a scotch and showing a boat, even if the skipper were black. And, as George Lois suggests, to them the Cutty Sark looks like a slave ship. Thus symbols and images are often totally different...
> Ads placing blacks in subservient positions received more negative responses from blacks; white responses were more neutral.
> (Gibson, 1978), pp. 60-84

Clearly when money matters, cultural sensitivity becomes an imperative. Businesses seem able to respond, why not educators?

Groups vary in the use of their environment. Individuals may behave in close harmony with the particular cultural group into which they were born. On the other hand a given individual or group may have learned patterns of other groups in addition to or in place of their own. Color alone may not be sufficient to identify a person as belonging to a "Black culture." Language alone is likewise insufficient to identify a person as belonging to a particular cultural group. Compare, for example, the culture of the majority of the Spanish-speaking Cubans with the majority of the Spanish-speaking Philippinos.

There is no easy road to cultural classification.

Much more could be added here. However, it should take little effort to see that every person or group of people will create things out of materials which are available to them at a particular place and time. Further, these creativities begin with the accumulated experience of a particular person and group.

Most aspects of culture for a given person or group are "invisible." (Hall, 1977) (Shuy, 1976) They are so fully learned and are so fully incorporated into daily living patterns that they seem to the members of the culture to be "normal." At times, it becomes hard for members of a given culture to accept the behavior of members of other cultures as "normal" or valid. Other cultures are visible only through one's own cultural "lenses," or "screens." (Nobles, 1976 b) Therefore, another culture cannot be comprehended or grasped fully because of the alien observers distorted perception (Hall, 1977) (Levi Strauss, 1966) (Cole and Scribner, 1974) (Ramirez and Cateneda, 1974) (Hilliard, 1976).

Jones (1963) gives us an excellent example with African and African-American music and its critics.

> The role of African music in the formulation of Afro-American music was misunderstood for a great many years. And the most obvious misunderstanding was one that perhaps only a Westerner would make, that African music "...although based on the same principles of European music, suffers from the African's lack of European technical skill in the fashioning of his crude instruments. Thus the strangeness and out-of-tune quality of a great many of the played notes." Musicologists of the eighteenth and nineteenth centuries, and even some from the twentieth, would speak of te "aberration" of the diatonic scale in African music. Or a man like Krehbiel could say: "There is a significance which I cannot fathom in the circumstance that the tones which seem rebellious to the negro's sense of intervallic property are the fourth and seventh of te diatonic major series and the fourth, sixth and seventh of the minor." Why did it not occur to him that perhaps the Africans were using not a diatonic scale, but an African scale, a scale that would seem ludicrous when analyzed by the normal methods of Western musicology? Even Ernest Borneman says: "It seems likely now that the common source of European and West African music was a simple non-hemitonic pentatone system. Although indigenous variants of the diatonic scale have been developed and preserved in Africa, modern West Africans who are not familiar with European music will tend to become uncertain when asked to sing in a tempered scale. This becomes particularly obvious when the third an seventh steps of a diatonic scale are approached. The singer almost invariably tries to skid around these steps with slides, slurs or vibrato effects so broad as to approach scalar value."
>
> These sliding and slurring effects in Afro-American music, the basic "aberrant" quality of a blues scale, are, of course, called "blueing" the notes. But why not of "scalar value?" It is my idea that this is a different scale.
>
> Sidney Finkelstein, in Jazz: A people's Music: "...these deviations from the pitch familiar to concert music are not, of course, the result of an inability to sing or play in tune. They mean that the blues are a non-diatonic music...
>
> pp. 24-25.

Jones goes on to interpret the distortion. It is an interpretation which is very familiar to many Americans who enjoy the experience of knowing more than one culture.

> There are still relatively cultivated Westerners who believe that before Giotto no one could reproduce the human figure well, or that the Egyptians painted their figures in profile because they could not do it any other way. The idea of progress, as it has infected all other areas of Western thought, is thus carried over into the arts as well. And so a Western listener will criticize the tonal and timbral qualities of an African or American Negro singer whose singing has a completely alien end as the "standard of excellence." The "hoarse, shrill" quality of African singers or of their cultural progeny, the blues singers, is thus attributed to their lack of proper vocal training, instead of to a conscious desire dictated by their own cultures to produce a prescribed and certainly calculated effect. A blues singer and, say a Wagnerian tenor cannot be compared to one another in any way. They issue from cultures that have almost nothing in common, and the musics they make are equally alien. The Western concept of "beauty" cannot be reconciled to African or Afro-American music (except perhaps now in the twentieth century, Afro-American music has enough of a Euro-American tradition to make it seem possible to judge it by purely Western standards. This is not quite true.) For a Westerner to say that the Wagnerian tenor's voice is "better" than the African singer's or the blues singer's is analogous to a non-Westerner disparaging Beethoven's Ninth Symphony because if wasn't improvised.
>
> pp. 29-30

To a student of music who is also a student of culture, the concept of cultural variation will be easy to grasp. To the student who is ignorant of culture, both his or her own culture and that of others will remain invisible or incomprehensible.

By way of further illustration of this, another quote from Mr. Borneman:

> "While the whole European tradition strives for regularity of pitch, of time, of timbre and of vibrato - the African tradition strives precisely for the negation of these elements. In language, the African tradition aims at circumlocution rather than at exact definition. The direct statement is considered crude and unimaginative; the veiling of all contents in ever-changing paraphrases is considered the criterion of intelligence and personality. In music, the same tendency towards obliquity and ellipsis is noticeable: no note is attacked straight; the voice or instrument always approaches it from above or below, plays around the implied pitch without ever remaining any length of time, and departs from it without ever having committed itself to a single meaning. The timbre is veiled and paraphrased by constantly changing vibrato, tremolo and overtone effects. The timing and accentuation, finally, are not stated, but implied or suggested. The denying or withholding of all signposts."

The rules in music are culturally specific. Ranking the cultural products would make no sense. A "norm" is a meaningful referent here only within a cultural system. The same principles apply to linguistic differences and to devices which depend upon language such as paper and pencil tests. Typical cross-cultural observations by culturally untrained observers results in the denial of data. It may also result in the error of interpreting the cultural substance of one group in terms of the cultural substance of another. (Schwaller de Lubicz, 1977) (Nobles, 1976.) The matter may become even more confused and confounded when we understand that members of two different cultural groups may, in a particular instance, exhibit virtually an identical overt behavior. Yet the meaning of that behavior can be different for both people.

Culture is real. It is represented by a particular group history and present configuration, and it can be ignored only at peril to the truth.

STANDARDIZATION IN THE FACE OF CULTURAL REALITY

The overwhelming majority of test and measurement professionals are not specialists in the study of culture, and are insensitive to gross sources of variation in all experimental settings. This is an academic failure which is reflected in three ways:

1. Among standardized test makers, there is a general ignorance of the literature about the investigator's own culture as a culture. (Hall, 1977) (Shuy, 1976).

2. Among standardized test makeres, there is a general ignorance of the literature which describes the culture of specific cultural group other than the investigator's own.

3. Among standardized test makers, there is a general ignorance of the literature which provides a metalanguage for communicating about cultures that are tested. (Hall, 1977) (Levi Strauss, 1966) (Labov, 1970) (Chomsky, 1957).

"... Considerable attention has been given to language. In this area, the deficit theory appears as the concept of "verbal deprivation": Negro children from the ghetto area receive little verbal stimulation, are said to hear very little well-formed language, and as a result are impoverished in their means of verbal expression: they cannot speak complete sentences, do not know the names of common objects, cannot form concepts or convey logical thoughts.

"Unfortunately, these notions are based upon the work of educational psychologists who know very little about language and even less about Negro children. The concept of verbal deprivation has no basis in social reality: in fact, Negro children in the urban ghettos receive a great deal of verbal stimulation, hear more well-formed sentences than middle-class children and participate fully in a highly verbal culture; they have the same basic vocabulary, possess the same capacity for conceptual learning, and use the same logic as any one else who learns to speak and understand English.

"The notion of "verbal deprivation" is a part of the most modern mythology of educational psychology, typical of the unfounded notions which tend to expand rapidly in our educational system. In past decades linguists have been as guilty as others in promoting such intellectual fashions at the expense of both teachers and children. But the myth of verbal deprivation is particularly dangerous, because it diverts the attention from real defects of our educational system to imaginary defects of the child; and as we shall see, it lead its sponsors inevitably to the hypothesis of the genetic inferiority of Negro children which it was originally designed to avoid...

"Linguists are also in an excellent position to assess Jensen's claim that the middle-class white population is superior to the working-class and Negro populations in the distribution of "Level II" or "conceptual" intelligence. The notion that large numbers of children have no capacity for conceptual thinking would inevitably mean that they speak a primitive language, for even the simplest linguistic rules we discussed above involve conceptual operations more complex than those used in the experiment cited by Jensen. Let us consider what is involved in the use of the general English rule that incorporates the negative with the first indefinite. To learn and use this rule, one must first identify the class of indefinites involved, any, one, ever which are formally quite diverse. How is this done? These indefinites share a number of common properties which can be expressed as the concepts 'indefinite', 'hypothetical' and 'non-partitive'. One might argue that these indefinites are learned as a simple list by "association" learning. But this is only one of the many syntactic rules involving indefinites - rules known to every speaker of English, which could not be learned except by an understanding of their common, abstract properties.

(Labov, 1970, pp. 153-186).

In this case the metalanguage would help the observer to focus on the logic of the discourse rather than upon the standardization of content. It would enable _false_ deficits to be correctly identified as such. It would also enable false superiority to be identified as such.

In every learned journal one can find examples of jargon and empty elaboration - and complaints about it. Is the "elaborated code" of Bernstein really so "flexible, detailed and subtle" as some psychologists believe? (Jensen 1968: 119). Isn't it also turgid, redundant, bombastic and empty? Is it not simply an elaborated style, rather than a superior code or system (10)

Our work in the speech community makes it painfully obvious that in many ways working-class speakers are more effective narrators, reasoners and debaters than many middle-class speakers who temporize, qualify, and lose their argument in a mass of irrelevant detail. Many academic writers try to rid themselves of that part of middle-class style that is empty pretension, and keep that part that is needed for precision. But the average middle-class speaker that we encounter makes no such effort; he is enmeshed in verbiage, the victim of sociolinguistic factors beyond his control.

(Labov, 1970, p. 164).

Those who are ignorant of the principles of culture tend to commit certain predictable errors.

1. They dismiss talk of culture and variation as matters of rhetoric, idealogy, politics or sentimentality. It must be noted that they do this summarily and without data.

2. They, then, proceed in attempting to force cultural realities into standadized, preconceived, a priori, categories or classifications.

One begins to suspect more politics than science here, especially among the professional disciplines. For it is clear that the power structure of the present standardized testing community would shift dramatically if the monopoly of the psychologists over school assessment were to be toppled. Cultural anthropology and sociolinguistics, among other academic disciplines, already have the tools to remedy this deplorable condition of cultural ignorance. (Shuy, Hall, Labov, Chomsky) But these and other relevant disciplines are virtually barred from the area of school assessment. Their knowledge base is virtually taboo. (Hilliard, 1979a) It appears that, among standardized test makers, the question of culture are not debated at all. Doing so is likely to result in one of three things:

1. A confession of ignorance of relevant empirical data.

2. A revelation of knowledge of relevant empirical data, but with a deliberate intent and calculation to conceal that knowledge in order to deceive audiences.

3. An adjustment of present assessment practice to accomodate empirical knowledge.

Consider the following example which illustrates the fundamental threat to "measurement" that is posed when standardized testing which relies on language is used. Linguists can illuminate clearly the folly of aggregating test item results where the linguistic meaning is variable among examinees. Roger Shuy (1976) has shown how learning to read is related to linguistic features which vary among diverse linguistic communities.

Practical experience indicates that different levels of language may take on prominence at different stages in the progression of reading skills. Thus, processing sound symbol correspondences may be relatively important for the reader in the beginning stages of reading, but they become less important as syntax and semantics become more important.

Phonology, morphology, lexicon, syntax and discourse are all culturally specific! Therefore, a standadized reading text or a standardized test of reading ability actually requires quite different things from various linguistic communities.

Aggregation, standardized scoring and comparison of performances where meanings vary is the height of scientific irresponsibility.

Margaret Donaldson illustrates this fallacy of a mismatch between the communication system or language of the examiner and that of the examinee.

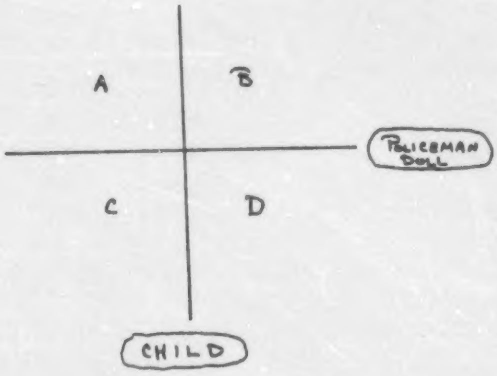
> We are urged by Piaget to believe that the child's behavior in this situation gives us a deep insight into the nature of his world. This world is held to be one that is composed largely of "false absolutes". That is to say, the child does not appreciate that what he sees is relative to his own position: he takes it to represent absolute truth or reality - the world as it really is. Notice that this implies a world marked by extreme discontinuity. Any change in position means abrupt change in the world and

a sharp break with the past. And indeed Piaget believes that this is how it is for the young child: that he lives in the state of the moment, not bothering himself with how things were just previously, with the relation of one state to those which come before or after it. His world is like a film run slowly, as Piaget says elsewhere.

This is by no means to say that Piaget thinks the child has no memory of the earlier "stills." The issue for Piaget is how the momentary states are linked, or fail to be linked, in the child's mind. The issue is how well the child can deal conceptually with the transitions between them.

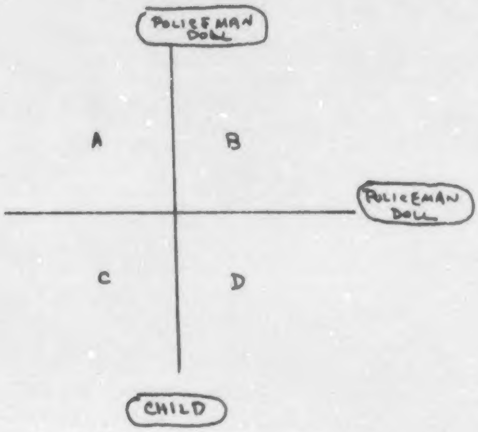
All this has far-reaching implications for the child's ability to think and reason, and we shall come back to these implications later. But first let us consider how children perform on a task which is in some ways very like the "mountains" task and in other extremely important ways very different.

This task was devised by Martin Hughes. In its simplest form, it makes use of two "walls" intersecting to form a cross, and two small dolls, representing respectively a policeman and a little boy....In the studies which Hughes conducted the policeman was placed initially as in the diagram so that he could see the areas marked B and D, while the areas A and C were hidden from him by the wall.



The child was then introduced to the task very carefully, in ways that were designed to give him every chance of understanding the situation fully and grasping what was being asked of him. First, Hughes put the boy doll in section A and asked if the policeman could see the boy there. The question was repeated for sections B, C, and D in turn. Next the policeman was placed on the opposite side, facing the wall that divides A from C, and the child was asked to "hide the doll so that the policeman can't see him." If the child made any mistakes at these preliminary stages, his error was pointed out to him, and the question was repeated until the correct answer was given. But very few mistakes were made.

Then the test proper began. And now the task was made more complex. Another policeman was produced and the two were positioned....(see Fig. 2).



> The child was told to hide the boy from both policemen, a result which could only be achieved by the consideration and coordination of two different points of view. This was repeated three times, so that each time a different section was left as the only hiding place.
>
> The results were dramatic. When thirty children between the ages of three-and-a-half and five years were given this task, 90 percent of their responses were correct. And even the ten youngest children, whose average age was only three years nine months, achieved a success rate of 88 percent.
>
> Hughes then went on to further trials, using more complex arrangements of walls, with as many as five or six sections, and introducing a third policeman. The three-year-olds had more trouble with this, but they still got over 60 percent of the trials correct. The four-year-olds could still succeed at the 90 percent level.
>
> Donaldson, 1978, pp 12-15)

In these and other studies cited by Donaldson, the critical role of language is clearly shown. Standardization is dependent upon a common language between examiners or tests and examinees. Donaldson is clear on this point.

> In any event, the questions the children were answering were frequently *not* the questions the experimenter had asked. The *children's* interpretations did not correspond to the *experimenter's* intention: nor could they be regarded as normal given the rules of the language. The children did not know what the experimenter meant: and one is tempted to say that they did not strictly appear to know what the language meant. Or, if that seems too strong, one must at least say that something other than "the rules of the language" was shaping their interpretation — something perhaps like an expectation about the question that would be asked, an expectation that could be influenced by the nature of the experimental material. However, it is essential to notice that we may not conclude that the children were, in some general way, not bothering to attend to the language — for we must recall the dramatic effect in some of the studies, of the inclusion or omission of a single adjective.
>
> (Donaldson, 1978, p 20)

It should take little imagination to see that if examiner and examinee are from the same culture and still misunderstand each other, the problem is exacerbated in cross-cultural settings. The cross-cultural setting is most frequent where African-American and Mexican-American children are concerned, since many if not most of the examiners and the tests which are used with these children are quite alien to their experience.

In short, if assessment, interrogation and interpretation becomes heavily dependent upon the surface features of a particular culture, systematic assessment can still survive. However, mass production of testing instruments, in this case, must be discontinued. Mass production of testing instruments will be appropriate for use with all cultural groups when those instruments are able to tap "deep structures."

WHY CULTURE IS USUALLY IGNORED IN TESTING AND ASSESSMENT

While culture is real and is a major variable in human experience, United States social science which supports standardized testing seems not to have caught on. There appear to be several reasons for this:

1. Culture is "invisible" or out of consciousness for most of those who have not been trained to perceive it.

2. The popular labels which are used to identify cultural groups are almost always confounded. They are not precisely defined terms. In fact, they are frequently undefined. For example:

    a. "Race" is not the same as culture. Therefore, terms such as "Caucasoid," "Negroid," or "Mongoloid," if they mean anything at all (Barzun, 1965) (Benedict, 1968) (Montagu, 1974) do not define cultural patterns.

    b. Geography is not equivalent to culture. Therefore, terms such as "German," "Asian," or "Mexican" do not define cultural patterns.

    c. Poverty is not equivalent to culture. Therefore, the condition of being without resources does not define a cultural pattern. Socioeconomic status (SES) is not cultural variable.

    d. Religion is not equivalent to culture. Therefore, terms such as "Jew," "Protestant," or "Catholic" do not define cultural patterns.

    e. "Minority" is not equivalent to culture. Therefore, the condition of being outnumbered does not define a cultural pattern.

3. A given cultural group may change its name. This may add to an observer's confusion. For example, are "Colored," "Negroes," "Blacks," or "African-Americans" the same in the United States of America? For example, does an alien observer determine who belongs in a "Black" sample for a cross cultural study?

4. The word "culture," while widely used, is not used with precision. That is to say, the dimensions of human behavior which may vary are not all kept in mind by those who use the word. Further, different users may give

different weights to the dimensions. For example, one person may say "culture" while meaning language and customs, such as dress and music. Another person say "culture," while meaning religious or philosophical beliefs or word view. There are some who use "culture" to mean the life styles of wealthy or powerful people in old families.

5. If culture is recognized, it introduces "inefficiency" in the testing processes which were intended to be universal. Therefore, there testing becomes more costly.

6. Since the unsophisticated identification of cultural groups tend to cause differences to be "washed out" in research, observers tend(to c nclude that culture is of little importance. For example, if "race" is equated to culture, different cultural groups may actually be grouped together, as if they were the same for research purposes.

7. The politics of a situation can function to shut down communication altogether. An observer of another person or group may see that person or group as a threat, or may have a vested interest in the exploitation of members of other cultures; then, the picture of the other culture which emerges will tend to be a self-serving rationalization of the alien groups cultural reality. (Pearce, 1965) (Stanton, 1960) (Weinreich, 1946) (Hilliard, 1979a).

CULTURE AS AN IMPERATIVE IN TESTING AND ASSESSMENT

We should be able to see at this point that all access to meaningful aspects of human behavior for instructional purposes is through culture. That is to say, whatever the underlying mental function or process which is being assessed (e.g., deep structures as in language (Chomsky, 1957) (Labov, 1970) it can only be manifest through the specific cultural material possessed by a learner. It is, therefore, a truism to say that no other option presents itself to us at this time.

THE MEANING OF "TEST"

The use of tests in education have become both arbitrary and ritualistic. The uses are said to be arbitrary because the link between "testing" and instructional improvement is seldom demonstrated. They are ritualistic in the sense that testing is an activity which most educators feel compelled to perform. Yet when asked why they do so, they become inarticulate, prone to the use of cliches, and tend to focus on irrelevant issues. (Hilliard, 1979b). These inadequate responses in education can be predicted readily if we compare the usual meaning of "test" in physics or chemistry, with the meaning of "test" in education. "Tests" in physics or chemistry or medicine are performed when the characteristic properties of measurement instruments are well known and when major sources of variation are controlled.

To "test" for the saline content of water is to know both the properties of water and various salts, as well as the nature of their interaction. To "test" blood for infection is to know the behavior of blood under healthy and sick conditions and how it may vary with specific types of infection. It should be clear that "test" in education is not nearly so precisely defined or employed. In general "tests" are poorly linked to professional practice. In fact, one should question whether the state of the art in testing or instruction is sufficiently developed and systematic to justify the use of the term "test" in its more traditional scientific sense. To qualify as a "test," accountability is required, over and beyond the simple criteria of instrument reliability and "predictive validity." Prediction is not explanation. Testing should contribute to explanation. If not, we should return to traditional achievement "examination." A true test should reveal with clarity some reality which would be obscure or ambiguous or invisible without the test. Perhaps the word examination should be reserved for inquiry which is designed to determine if certain skills and content are present. Then the term test could be used for those systematic inquiries which are designed to render information which explains teaching and learning processes. Examinations can tell us what. Tests should tell us why. In any event, there are two very different functions which assessors perform that require quite different designations, if confusion is to be avoided.

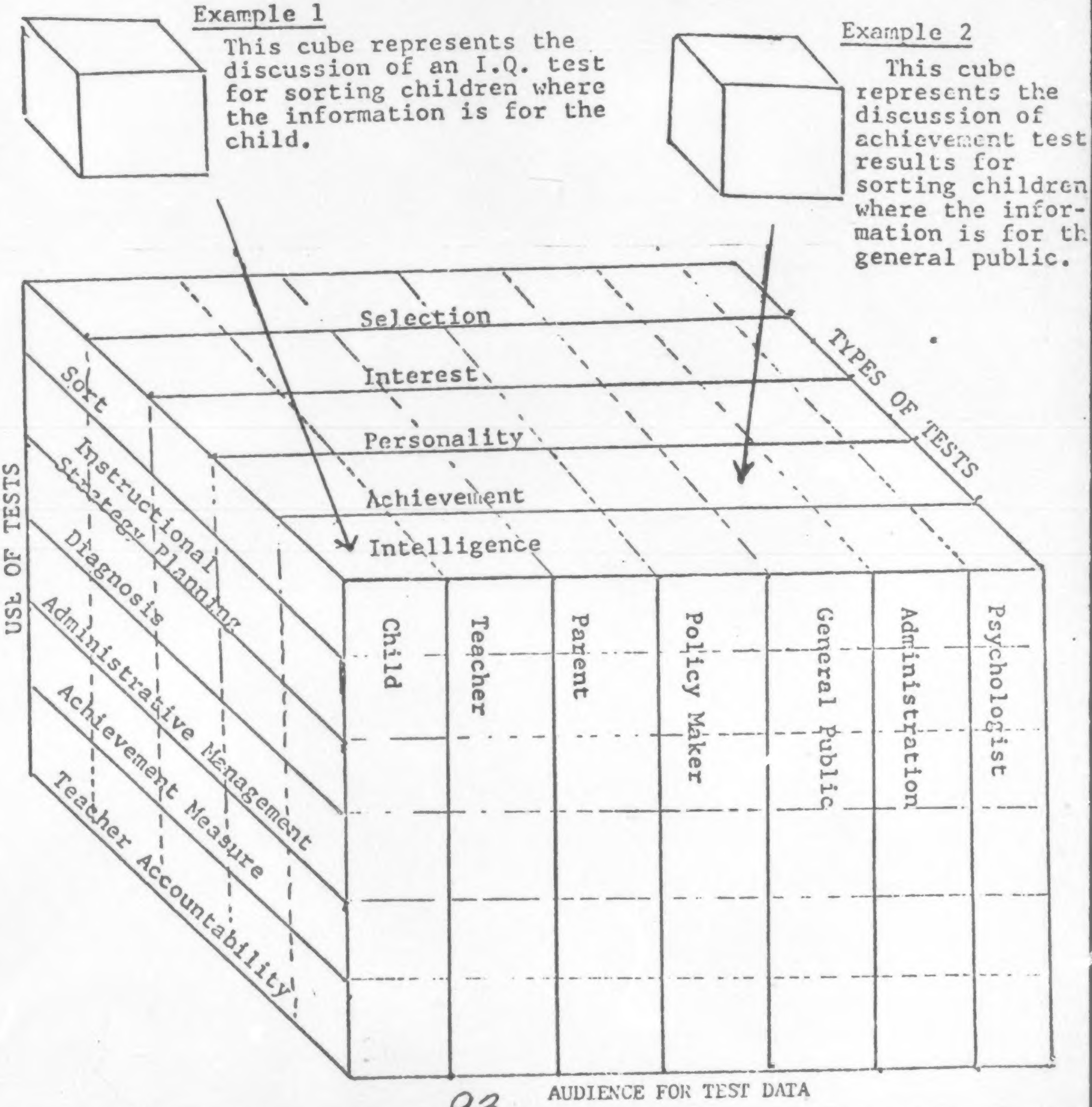
IT IS IMPORTANT TO EXPRESS CERTAIN IMPLICIT ASSUMPTIONS UPON WHICH SYSTEMATIC ASSESSMENT IS BASED. THESE HOLD ESPECIALLY FOR STANDARDIZED TESTING.

> SYSTEMATIC ASSESSMENT WILL ENABLE COMPARISONS TO BE MADE AMONG INDIVIDUALS AND GROUPS. THESE COMPARISONS ARE BASED UPON A CRITERION OR CRITERIA WHICH HAVE STABLE MEANING ACROSS INDIVIDUALS AND OR GROUPS.
>
> SYSTEMATIC ASSESSMENT IN EDUCATION IS EQUIVALENT TO MEASUREMENT IN THE PHYSICAL SCIENCES.
>
> THE AGGREGATION OF SCORES ON PAPER AND PENCIL TESTS IS THE SAME AS THE AGGREGATION OF COMPERABLE UNITS OF BEHAVIOR. EACH ITEM IS SEEN AS BEING THE SAME AMOUNT OF THE SAME KIND OF BEHAVIOR. IF THEY ARE NOT REGARDED IN THIS WAY, THE AGGREGATION IS INAPPROPRIATE.
>
> TEST ITEMS SHOULD HAVE UNIQUE RIGHT ANSWERS.
>
> TEST DATA GUIDE ISTRUCTIONAL STRATEGY VALIDLY, I.E., INSTRUCTION WILL BE BETTER BECAUSE OF THE USE OF ASSESSMENT.
>
> TEST ITEMS SAMPLE ADEQUATELY THE DOMAIN WHICH IS BEING EXAMINED.
>
> VALID INSTRUCTIONAL STRATEGIES EXIST WHICH REQUIRE ASSESSMENT DATA IN ORDER TO BE EMPLOYED.

IN GENERAL, THESE ASSUMPTIONS CANNOT BE MET IN PRACTICE. STATISTICAL PROCEDURES FOR PROCESSING "DATA" ARE HIGHLY SOPHISTICATED. YET, "DATA" FOR PROCESSING ARE FREQUENTLY MISSING OR CONFOUNDED. IF THESE ASSUMPTIONS CANNOT BE MET, RITUAL AND SUPERSTITION WILL CONTINUE TO PREVAIL.

A PARADIGM FOR SORTING DISCUSSION ISSUES IN TESTING

An abundance of issues in testing are frequently lumped together in discussions and analyses. These must be sorted out and discussed one at a time, if clarity is to obtain. When speaking of testing, it is important that the use of test, the audience or user of the information and the type of test being used be identified. Discussants must, in general, talk in one of the cells of the paradigm at a time, or discourse will be confounded.



Example 1

This cube represents the discussion of an I.Q. test for sorting children where the information is for the child.

Example 2

This cube represents the discussion of achievement test results for sorting children where the information is for th general public.

Selection
Interest
Personality
Achievement
Intelligence
TYPES OF TESTS
USE OF TESTS
Sort
Instructional Strategy Planning
Diagnosis
Administrative Management
Achievement Measure
Teacher Accountability
Child
Teacher
Parent
Policy Maker
General Public
Administration
Psychologist
AUDIENCE FOR TEST DATA

The culturally sensitive use of tests will depend upon clarity about the uses of assessment and the user audience. For example, I.Q. tests are used as sorting instruments by administrators. Yet the myth prevails that they also are used as diagnostic devices for the development of instructional strategies. The cultural bias of the I.Q. test is justified by some who argue as if the I.Q. test were really an achievement test for all audiences, i.e. that mainstream culture and the schools require a certain culture's vocabulary and tha certain types of problems be solved. That argument represents a shift from thinking of the I.Q. test as a mental measurement device to an achievement measurement device. The shift is a major one which fuses two types of uses unsystematically into one discussion.

Among the current uses of testing and assessment, the most easily justified is the assessment of achievement. There the major issues for particular cultural groups on this type of test are content validity and communication accuracy. On the other hand, where I.Q. testing is involved, the major issues for particular cultural groups are construct validity, as well as diagnostic validity. Here, there are quite general grounds for questioning the validity of the instruments (see for example Houts 1977) Indeed, the I.Q. score is a worthless piece of information. (Hilliard, 1979b)

There are many ways in which the consideration of audiences for and uses of tests can bring clarity to a heavily confunded area. When tests are used across cultures, such a paradigm becomes an imperative.

PRINCIPLES FOR CULTURALLY SENSITIVE ASSESSMENT

Any test or assessment procedure which responds to and uses the culture of students would follow certain principles. I believe that the following would result in greater validity for assessment.

1. Test and assessment procedures must reflect consciously a sensitivity to the unique culture of the learner. Ernie A. Smith illustrates in a precise way how culture (language) specific tests might be constructed. By use of such culture specific tests, it should be possible to determine if a child is speaking and hearing in harmony with a specific linguistic community. If so, the a speech assessor would rule out pathology, a reading teacher would rule out "reading error," and a psychologist would rule out mental deficiency when basing such judgements on the childs "dropping" of final consonant clusters.

2. Tests and assessment procedures must yield a description of the learner's repetoire, not simply the presence or absence of material from the test makers repetoire. (Lee C. Lee, 1978)

3. Tests and assessment procedures must yield a description of learners processes, not simply the content of responses to questions. (Piaget, 1970) (Brown and Burton, 1979)

4. Tests and assessment procedures must yield a description of the learners progress, and not simply the learners status at a particular time. To be meaningful the assessment of progress must be accompanied by a description of the teaching services which were provided to the student. (Kunzleman and Koenig)

5. Tests and assessment procedures must yield a description of the teacher/learner and or tester/leaner interaction, since teachers and testers are non-student sources of variation. (Rist, 1973) (Hamilton, 1974)

6. Tests and assessment procedures must yield a description of the general ecology of the testing setting.

7. Tests and assessment procedures must be related clearly to a valid theory of healthy or pathological functioning and valid professional intervention.

Among other aims, systematic testing and assessment in education should be used to assess changes in learners, to reveal patterns of learner behavior and to guide teaching strategies. New insight and effective communication should result from good assessment.

FOSTERING INTELLECTUAL DEVELOPMENT

Howard E. Gruber
Robert T. Keegan

Institute for Cognitive Studies
Rutgers University
Newark, NJ 07102

In this paper I want to describe a method of teaching that grew out of a joint concern for one product of the evaluation industry and one aspect of teaching. The story has three strands. First strand: When Arthur Jensen published his famous Harvard Educational Review paper, apart from certain technical criticisms, my main reaction was to dream of a demonstration experiment that would show clearly that intellectual functioning could be drastically modified by changes in educational practice. This would make nonsense of the heritability argument as Jensen used it. Dobzhansky (1972) has since written an excellent critique of this usage of the concept, but he bases his theoretical argument on results obtained from experiments with fruit flies. My thinking ran along similar lines, but I was interested in human experience.

Second strand: My children attended the Free School of Bergen County, a high school outside the public school system, run by the children themselves, that had a noticeably longer life than many similar ventures. I taught there a little. On the first day I arrived with a carefully planned lecture on imagery, a topic that forms part of my research interests and that I know interests most people. There were 20 students from 13-17 years old, sprawled on a rug and sitting on cast-off couches. I chose a vacant spot — on a piano stool by an old piano. Looking around, I wondered if my lecture was a good way to begin. Entirely on impulse, I told them in a sentence or two about synesthesia — sometimes an auditory stimulus elicits a visual experience (or other such combinations). Then I asked them to close their eyes and try to see something when I played a note. We went around the room, each person describing what he or she saw. Both the diversity and the commonalities were intriguing. The class's attention engaged, I drew breath and was once more about to start my lecture. Someone called out, "let's do that again!" I complied. We repeated, over and over, with many variations. New facets of a complex process emerged— Try to imagine a pure color, or a scene with motion, etc. This time, no sound, imagine your breakfast table (shades of Francis Galton), etc. Very occasionally, I made a remark about psychologists' previous work on visual imagery. Suddenly our time was up — 1-1/2 hours have flown by. My lecture had become irrelevant. Actually, all the essentials came up one way or another in our explorations. The students had discovered almost everything, and interest had not flagged for an instant. (Later, when I thought about the relation between what we had done and the "discovery method" it struck me that one important difference was that I had had no set objective, nothing in particular I thought it was essential for them to discover.) I came out of the school flying high and wondering, "why can't college teaching be as exhilirating as that?"

Third strand: Over a long period of time, I received inquiries from the Academic Foundations Department at Rutgers University-Newark (a Seek-type program): what are cognitive skills? how can we develop them? I resisted giving any positive response to these inquiries, because I felt that the Department was terribly addicted to using workbooks and other mechanical devices, taking a routine and Skinnerizing attitude toward the question of remediation. Eventually, I began to feel that my attitude ought to be more constructive, and I began to wonder what we, as an Institute for Cognitive Studies, could do. My general approach was this: The human mind is a wonderful instrument. When it is working well, people can do what they want, learn what they want and need. When it is not working well, all the workbooks in the world won't help. Question — How do you get people to think better? I put this question to a graduate seminar as our term project. Their first reaction was to raise an ethical protest: Who were we to tell other people how to think? Struggling with this problem had an important and I think profoundly beneficial effect on the program we worked out.

ON TEACHING PEOPLE TO THINK BETTER

I shall talk about the need for and possibility of educational programs which make direct attempts to teach people to think better. Most of what I have to say will deal with the need, rather than with a detailed description of actual programs we have conducted in several settings. Our work is only a few years old; we are still in the inventing phase, and not yet ready to proclaim our methods from the housetops.

The aim of the program we have been developing in Newark is to develop a method for teaching people to think better. While our primary goal is in the field of innovative teaching methods, a fundamental psychological question is also at stake: can we alter the course of intellectual growth after the early formative years of childhood?

I begin with three examples to show that the acquisition of verbal and symbolic skills in a conventional way, even to an exceptional degree, do not necessarily indicate equally satisfying intellectual functioning. At the level of professional life, the Soviet psychologist Luria described the now celebrated case of Mr. S, who was gifted with extraordinary powers of visualisation and memory, was not in other respects a particularly gifted person, and in some ways he was rather limited. At the level of graduate school pereformance, it is now notorious that high scores on the Graduate Record Examination, emphasizing verbal and quantitative skills, correlate very poorly with success in graduate school. At the level of undergraduate performance, Professor David Griffiths of Essex County College has recently shown that students receiving a grade of C or better in college level introductory physics and chemistry courses have often achieved this success without being able to reason at the level of formal operations as described by Jean Piaget. What is perhaps more significant in the present context, Griffiths found that, in a typical state university population, there were many students who seemed to have succeeded in their science courses on the basis of a thin veneer of verbal skills, without any general or abstract grasp of what they were learning; meanwhile, at a nearby community colllege, the students performed as well — or as badly, if you prefer — on tests of formal reasoning, but lacked the aforementioned verbal veneer. Finally, as might be expected, there seemed to be at least some cases in which this verbal veneer actually got in the way of good thinking. Someone once said, "Words are a writer's worst enemy." The rest of us should watch out for them too.

> Insofar as "evaluation" is meant to measure retention of what has been learned, at least some of it should be done long after the learning experience. That is the way to find out how well education works. But there is very little research on really long-term retention, and most teachers have no information at all about the abiding consequences of their work. The hard truth must someday be faced that long-term retention is bound up with understanding. And understanding cannot be achieved by skipping quickly through 30 topics, 2 per week, as in many a typical course. An anecdote may help to bring out these points:
>
> One of the present authors has long been interested in people's understanding of elementary physical principles. In a relaxed and pleasant setting at a summer place, by way of explaining this part of my work to an old friend, I asked her a simple question — "Suppose you are in a closed railway car traveling in a straight line at constant speed. You stick out your hand in the aisle and drop a tennis ball. Where does it land?" After a long silence, she bursts into tears and sobbing. "Margot! Why are you crying?" "Because I got an A in college physics!" From a strong, traditional university, I might add.
>
> This refrain, "I did well in the course, but I never understood anything." -- a remark Margot made later — arose also in a series of interviews on physical knowledge and understanding conducted by Professor Andrea di Sessa and me at M.I.T.

I do not mention these facts in order to denigrate or minimize the importance of organized knowledge and fundamental symbolic skills such as reading. But it is vital to see the process by which such knowledge and skills are mastered in its total psychological and intellectual context.

Imagine a hypothetical case, for example, an individual with a miserable high school background, now in his or her middle twenties, finds his way back to school in a community college setting. He recognizes some fundamental deficiencies in his academic skills and wants to correct them. Now let me add one further premise — that this individual is what we really mean by a good student -- someone going to college to improve his mind, to have a rewarding experience of personal intellectual growth.

For awhile, our hypothetical student may be cajoled or coerced into various training programs narrowly focussed on particular problems of remediation. But he is too mature and sophisticated to be very long seduced by the allure of getting good grades, and he is skeptical about any promises of a relation between grades and eventual success in climbing some career ladder. The good student wants the remedial work if and only if it is clearly a part of a rewarding experience of personal intellectual growth.

Our hypothetical student may not be aware of all these subtleties at the outset of his post-secondary education, but as he progresses in his walk through the groves of Academe, he becomes increasingly aware of the fungus on some of the trees, the dead wood, and the stagnant pools across paths going nowhere. With a little more luck he may also become aware of, or begin to dream of, another part of the forest, where things are growing better. We may flatter ourselves into thinking that we can keep him pointed toward his workbooks and away from the Tree of Higher Knowledge...But the secret gets out. He has his own ways of knowing that when he is all done with his remediation things may be no better for him than they were for Luria's Mr. S. or for Griffiths' subjects.

For all we know, it may well be the better souls — the more perceptive, honest, and hardier individuals — who have the courage to recognize these potential limitations of the college experience and to turn away from it. It may be the good student — the one seeking a rewarding experience of personal intellectual growth — who most clearly sees the main danger of college life: not only does it disappoint, it may even spoil your mind.

Some of you who are extreme skeptics may think that good students such as the one I have imagined are rare and vanishing. But in our Practicum for the Improvement of Cognitive Functioning, where such issues are brought out into the open, we find this kind of good student to be the rule, not the exception. And Hans Furth, and Harry Ward in their book, THINKING GOES TO SCHOOL -- PIAGET'S THEORY IN PRACTICE, were describing the attitudes of the child who would become this young adult when they wrote, "The permutation game is a developmentally high-level activity which carries its intrinsic motivation whereas reading is often experienced as low-level activity." (p. 271)

If you struggle with the question. "How do you get people to think better?", a number of reasonable responses come to mind. The plurality of these responses should not be viewed as a problem but rather as indicative of the fact that "good" thinking is not a monolithic process. "Good" thinking is productive thinking and it requires many complementary component skills. In the attempt to develop an effective program for improving the quality of an individual's thought processes, several themes forcefully emerge. These central themes serve as a guide for the development of the particular tasks or "situations" that we utilize in the classroom. One of these themes has already been alluded to, namely, the great advantage of having access to large array of cognitive skills. A large array allows for flexibility of thought or, stated a litle differently, it enables an individual to have more than one way of thinking about whatever he wants to think about. Repetoire enlargement then, is one of the central themes to be discussed.

The classroom setting is well suited to the task of enlarging a student's repetoire of cognitive skills because it contains diversity, a key element in effectuating this expansion. Each student has his own way of approaching a task and in many cases, the student feels that his particular approach is the only conceivable method of operation. However, the inevitable diversity of approaches among individuals in a classroom provides a rich natural resource for exploration. In order to take advantage of this pool of diverse responses, the teacher has to assume the role of a moderator, rather than the more traditional "lecturer" role. The teacher focuses the dialogue among the students, emphasizes certain points, and does some degree of synthesis, but the "food for thought" arises from the students' interaction with each other. Dialogue offeres an individual the opportunity to see his own thought processes and capabilities mirrored in others. Feelings of, "I never thought of it in that way before" or "That's where I was going wrong", are compelling learning experiences.

We see then that the expression of diverse approaches to a situation or problem can expand the repetoire of cognitive skills of an individual by making him aware of approaches or strategies which had never before been available to him. But this is not enough. Outside the classroom, when an individual is confronted with a puzzling situation calling for a novel response, the individual may not have access to a group of people with whom he can enter into a dialogue for the purpose of expanding his range of possible responses. The individual must be able to generate alternatives on his own. The model of external dialogue has to be internalized for it to be of lasting worth. This internal dialogue is an intrinsic part of a reflective cognitive style.

Reflectivity is another of the central themes concerning productive thinking we referred to above. Reflectivity is not one particular cognitive skill, but rather a constellation of skills that can be thought of as constituting a cognitive style. Promoting this style is a cornerstone of our approach to education.

We have already described reflectivity as having the flavor of an internal dialogue. This description, however, should not be taken literally to mean that a fully composed subvocal conversation takes place in the head. As in a normal dialogue, the essence of understanding derives from the attempt to reconstruct the point of view of the other, the effort to attend to what the other is saying, and a certain degree of "reading between the lines". The following section deals with the role of "point of view", "paying attention" and "reading between the lines" in the internal dialogue characteristic of the reflective style.

The term "dialogue" presupposes more than one point of view and for this reason it captures the essence of the reflective style much more accurately than conceptualizing reflectivity as consisting of an internal monologue. In our discussion of the theme of the expanded repetoire, we pointed to the efficacy of the dialogic process in bringing about an experience of; "I never thought of it in that way before". The construction of a new point of view, which often consists of restructuring familiar material, constitutes an expansion of the individual's repetoire. Moreover, experiences of this kind should eventually communicate the point to the individual that the particular approach that first comes to mind is not the only approach possible. This realization, in and of itself, is a surprising revelation for many. The function of internal dialogue is not so much to denote the "correct solution" to a situation but to supply the right questions, questions such as, "Is there another way of approaching the situation?", or "How can I change the situation in order to clarify it?" These types of questions can serve as guidelines for constructing a new point of view.

It may at first appear trite to say that "paying attention" can help understanding, but "paying attention" has a special meaning with respect to reflectivity. All through childhood we are told to pay attention to our parents, our teachers, and to numerous others who have some degree of authority, but we are seldom if ever told to pay attention to ourselves. Reflectivity requires an individual to pay attention to his "line of thought". By closely attending to your own thought processes a whole set of experiences that was felt to be of little use now assumes great importance. The type of experiences we refer to is the commission of errors.

Piaget has shown us the value of attending to errors in the analysis of children's thinking. A careful analysis of the protocols of children who commit errors on a particular task can reveal information about the structure of the child's thought which cannot be determined by looking at "correct" or successful performance. Why shouldn't the same be true at the adult level? Errors can provide a wellspring of information for the attentive cognizer. For one thing, errors can be used to isolate problem areas in a line of thought. In striving to understand any particular unknown there ar usually various nexuses at which certain decisions and assumptions must be made in order to carry forward the cognitive task at hand. When the cognizer gains knowledge of an error he can either passively accept the correction and proceed accordngly or he can use the experience to identify the point at which he went wrong. In and of itself, commission of an error is not a valuable tool for learning but it can function as such if the learner takes the opportunity to discover the discrepancy between his chosen approach and other possible approaches. Reflectivity involves periodically "rewinding" the stream of thought, identifying the points of conflict or ambiguity where the error arose, and generating alternative courses of action (points of view).

Earlier we stated that understanding a dialogue entailed a certain degree of "reading between the lines" and that the same process was true for the internal dialogue characteristic of the reflective style. When you "read between the lines" you glean information that is not explicitly present but is somehow hidden and implied. The transformation of implicit knowledge into explicit knowledge is an act of creative liberation. Explicit knowledge can be used in a manner in which implicit knowledge cannot. Explicit knowledge is definable, mobile, and versatile by way of these characteristics. Implicit knowledge, on the other hand, is amorphous, frozen and non-versatile because it is so deeply embedded in context. Where knowledge that is explicit can be utilized in numerous contexts, implicit knowledge cannot be utilized in such a way because it has not yet been differentiated from context, and therefore it is confined to play a limited and limiting role in the cognitive life of the individual.

The recognition of the role that implicit knowledge plays in cognitive life is reflected in the concept of the presupposition. In even the simplest statement, numerous "unconscious" assumptions are made. For instance, in the simple request to "Please pass the sugar", the speaker assumes that the person spoken to can understand English, is physically capable of carrying out the act, is socially inclined to cooperation, and can identify sugar whether it be contained in a bowl, packet, or cube. Of course, each of these assumptions can be further divided into additional assumptions. The supposition that a person "understands English" involves assumptions concerning lexicon, syntax, phonology, contextual meaning, etc. While presuppositions are certainly a necessary component in "economizing" cognitive life, unconscious assumptions can also prevent productive thought from occurring. A prime example of how an assumption can debilitate thought occurs in the form of the syllogism. In order to correct a syllogistic line of reasoning it is necessary to specify the premises on which the reasoning is based and to root out the tacit implications contained in the premises. The act of making the implicit assumptions explicit liberates one from the fallacious line of reasoning.

Implicit knowledge forcefully affects the construction of a point of view, and a point of view serves as a guide to action and further thought about a subject. If I assume the world is flat then I will not attempt to sail "around" the earth. If someone else tries to sail around the earth and they do not return, I will then confidently conclude that the foolhardy crew fell "over the edge". Facts get interpreted in a manner that makes them consistent with the harbored point of view.

A recognition of the potent role of implicit knowledge in cognitive life and the intention to root out the presuppositions of a line of thought are powereful tools in achieving the development of a reflective cognitive style. A modicum of playfulness or "mischief" can be helpful in developing this reflective style. Assuming the role of a "devil's advocate" can also be instructive. For instance, if you start with the assumption "the earth is at the center of the universe", what does this assumption do to your thinking about all the other celestial bodies? This type of game playing can be quite instructive and liberating in that it helps to specify the place of assumptions in a line of thought.

From the preceding description of the goals of our course it should be evident that these goals are not specific to a course in Introductory Psychology. The skills outlined here are applicable in a wide variety of contexts, and that is precisely why they are of significant value. Later, we hope to show that not only the goals, but also the method we utilize is compatible with the teaching of other disciplines. The specific subject matter we worked with should not obscure the versatility of the method or the validity of the underlying goals.

ON SAWING A BOARD IN HALF, THINKING AND SELF-CRITICISM

In conventional education the functions of instruction and evaluation are kept separate. First, the teacher teaches and the students learn. Then, the teacher tests, then the students show what they have learned, and then the teacher evaluates this performance. In conventional education it is not thought bizarre to separate the person still further from the process of evaluation. The "test" may be taken away from the student, sent to another city, put into a machine, and transformed into number that bears absolutely no resemblance to what the learner learned.

Not all human activity is organized in this way. In some instances, performace and evaluation are inseparable. The carpenter rules a line and uses it to saw a board to a desired length. Every stroke of the saw is guided by the line and by the immediately visible pereformance. Corrections are not made because a third party ordains it, but in the dignified transaction between the sawyer and his work. There is a vital correspondence between the "test" administered by the ruled line and the work being done.

Let us examine for a moment these two educational structures; that arranged for the communication of knowledge and that ordained for the evaluation of the student's success in playing an appointed role in the communicative process. To simplify, we will consider only one type of class, a typical lecture course.

First we look at the structure of communication. In a typical lecture, the main activity can be described as one-many and one-way: one teacher talks to many students, and communication is almost entirely from teacher to students. Even a sensitive and concerned teacher has little opportunity to know what the students are thinking. They are silent. After class, it would be unusual for the teacher to look at some students' notebooks.

The teacher prepares carefully, works hard and continuously in class. But he or she works largely in ignorance of what the students are thinking meantime. Of course, we teachers tell ourselves about non-verbal communication, facial expressions, and occasional questions -- but all this gives only a very blurred reflection of the richness of our knowledge and thought we exhibit for our students.

To find out how much they have absorbed, we must typically wait until it is time for the test. Unlike the sawyer at work, the test is usually considerably separate in time from the rest of the activity. Moreover, the student does not evaluate his own performance. Often enough, he has only a vague idea of the criteria the teacher has used. Even when the teacher tries to spell these out, this explanation is seen as part of the structure of evaluation. Time spent in it is therefore time stolen from the more important structure of communication. Finally, the test result is given back to the student at a still later time, and transformed into a number. This number makes all sorts of administrative acts possible. But it does not convey at all the teacher's impressions of what the student actually did.

Some of the consequences of this arrangement:

1. Time taken for evaluation is minimized because it is seen as alien to the communication of knowledge.

2. Although the teacher hopes the students will focus their attention on the whole of what is being communicated, the students are using their ingenuity to figure out what fragments will be evaluated.

> Teacher (at end of inspiring lecture): Any questions?
>
> Student (raising hand eagerly): Will that be on the test, sir?

3. Little attention is given to providing the student with internal criteria for self-evaluation. The student lives in a world where, for many formative years, communication, performance, and evaluation are kept separate, and where some one else has the responsibility for evaluating the work done.

Now let us suppose that we take seriously the educational goal of helping students to become (or perhaps simply to remain) independent human beings, interested in and capable of evaluating their own performances.

What are some of the things a professional worker does? First, he or she has internalized criteria and a continuous sense of whether or not the work is going well. Second, since criteria are not so easy to come by, he or she spends a good deal of time developing them -- talking with colleagues, reading a critical literature, reflecting on his or her aims and progress. Third, when "outside" evaluation is needed, the professional thinks about whose opinion might be helpful and seeks it out. Fourth, this opinion is not sought in order to put an alien number in a record book. The worker wants criticism that corresponds to, is germane to, captures something of the work itself; such critical commentary is often a re-description of the work. And finally, the criticism is not merely listened to and the work then put away. The worker alters the work in some way that is responsive to the criticism.

It is a striking fact than none of these attributes characterize the main evaluation processes used in formal education. Conclusion: we are not teaching our students to be independent, self-evaluating human beings.

Our explicit goal in this project has always been "to help students to think better." Almost inadvertently, we have found that we have also been exploring new ways of coordinating the structures of communication and evaluation — and ways of modifying the process of evaluation itself to encourage the students to become people capable of self-criticism.

This new perspective on evaluation was a by-product of our work.(1) But now that it is there, it seems obvious that self-criticism is an integral part of good thinking and should always have been one of our goals. Sometimes making goals explicit facilitates pursuing them. Our work will probably change now, with this new-found recognition.

METHOD

In our attempt to help our students to think better, we have had to depart from the lecture as the primary means of "educating" our students. Instead, we have utilized three alternative classroom structures or formats that substantially modify both the students' and the teacher's conventional roles. Each of these formats has its distinct advantages, but, in the main, they all require a student to actively participate in a set task and to reflect upon his own course of action. By the same token, these alternatives also require restraint and patience on the part of the teacher. The student must be allowed to pursue his own method of dealing with the task, to make his own mistakes, and to develop his particular line of thought free from the well intentioned but ill timed intervention on the teacher. By allowing the student to discover the subject matter for himself, you largely obviate the necessity of lecturing to him about it, and class time can be used in more flexible and productive ways.

The "Round Robin" format has the great advantage of ensuring the participation of everyone in the class. What typically happens in this format is that a task is described to the class and everyone is asked to work on it. Following a period of individual work on the task, every member of the class is asked in turn to report on his experience in dealing with the task. It is essential that every student be heard from and that none be allowed to withdraw from the process. It is important not only for the student himself to become acquainted with this method of analyzing his own thought processes but also for the other students to have a point of comparison for their experiences with the same task.

The round robin format is not without its drawbacks. Inevitably, several students in any group will say things such as, "I did it the same way as John did it" when asked to give their reports. Replies of this sort do not have to be accepted. The teacher can carry forward the discussion by asking the student to describe in what way he felt his experience to be the same, in what wa it might have differed, or simply to put the experience into his own words. It is surprising how often this technique will uncover some new slant on the material. It also prevents other students from attempting to withdraw from the situation through this "me too" technique.

Another aspect of this round robin structure that may appear to be a drawback involves the time element. It takes time to go around the room and listen to the report of each student with care and interest. There is no quick way to explore the diversity of responses, identify the common elements, and emphasize the productive aspects of the material that arises from a class. This restriction becomes more salient with large class sizes. One way of dealing with this situation is to use the round robin in conjunction with other formats that also promote a reflective cognitive style.

The "Small Group" format also involves the initial presentation of a problelm or task, but with this format the active work on the task occurs in groups of four or five. The members of the groups are encouraged to freely exchange their viewpoints, offer tentative solutions, and work toward successful completion of task. Working in a small group provides quite a different experience from an individual's encounter with a task. Interestingly, what is often hidden from ourselves is clearly revealed in others. The round robin format makes it apparent that describing one's own thinking is a difficult task, but some of the same students who find the self report so difficult can facilely describe how someone else in the group handled the task, and how the group as a whole proceeded. Over time, the experience of attending to and describing the group process, as well as receiving feedback on his own contributions to the group, should help the individual to develop an analytical sense of his own thinking. When it comes time for hearing the reports on the progress of the groups, this can be handled by having one, several, or all the members of the group give their descriptions, effectively creating a round robin type of situation. There is nothing pure or sacrosanct about these formats and they can be used in interesting combinations.

The distinguishing characteristic of the "Demonstration" format is the departure from assigning the identical task to everyone in the class. From time to time we have found it useful to split the class into a few large groups and assign each group a slightly different variation of a task to perform. This structuring of the class enables us to single out specific factors (i.e. organization) and examine their influence on the way we think. Although the chief purpose of the exercise is to provide a clear and compelling demonstration of the selected factor at work, the differences between individual approaches to the task can and should also be examined. The round robin can be utilized to accomplish this aspect of the exercise.

On occasion, we have used a "mini-lecture" at the end of a class session to clarify a point or extend a topic in a new direction. Although this is a small concession to the conventional classroom model, we justify its use by keeping the number of such mini-lectures at a minimum, placing them at the end of the session, and by making sure that each mini-lecture is of a short duration. However, it is the round robin, small group, and demonstration formats that constitute the essence of our approach. The examples that follow provide a good picture of these formats in action.

"POWERFUL IDEAS", THINKING, AND REPRESENTATION

At the first meeting of the class in September 1979, we wanted to try introducing the course in a new way. Often, we begin by moving the students immediately into the round-robin format, demonstrate the process at work, and then discuss the reasons for it. This time we wanted to see if the students could play some part in generating the rationale for the course.

Leaving the students in the ordinary classroom formation, row behind row, I began by pointing out that there are many subject-matter options open to the teacher of introductory psychology, or for that matter, any course. How to choose?

Discussion was lively. We rushed the pace a little and as soon as it became possible seized on a student's remark to steer the discussion toward expressing the strategy of choosing "powerful ideas." I have heard much mention of this phrase in the last year or two, but no clear idea given as to how to know when an idea is powerful. My own guess is that no idea is of itself powerful. A person makes an idea powerful by linking it with other ideas. Some of this came out in discussion. Fairly rapidly we got around to the idea that the most valuable thing that could happen to a person in school would be to learn to think better, or to use his or her mind better. At this point, I explained that that was indeed the primary goal of the course, and introduced the first exercise, which really involved three steps.

First, I asked them to write down a paragraph or so explaining what they meant by "thinking." Second — after some discussion of "representation" as a powerful idea — I asked them to draw two diagrams, a diagram of a dialogue and a diagram of an ordinary classroom situation, depicting the pattern of communication among the paticipants. The paragraph on thinking was taken up in a later class meeting, where we worked out a way of categorizing the different responses, and discussed the class's ideas. The diagrams of dialogues were mainly of two persons linked by arrows or lines, indicating that they were talking to each other. In what follows we examine only the students' diagrams of an ordinary classroom situation.

These diagrams were divided into the six categories shown in Table 1. It can be seen that by far the dominant description is a one-way, one-many interaction between the one teacher and the many students.

TABLE 1

CATEGORY	
A. One-way, one-many communication (one teacher telling many students)	17
B. As above, but with some side-chains, (one teacher telling many students, some interstudent communication)	12
C. As above, but with every student engaged in some interstudent communication	1
D. Teacher-student interaction seen as a set of 1-1 dialogues	1
E. Two phases: A above, and ring structure indicating many-student discussion	1
Unclassifiable	4

(1) We were probably only dimly aware of the point — not yet having clearly delineated the separate but corresponding structures of communication and evaluation. At the March 1979 meeting of the panel on Assessment in the Service of Instruction, one of the authors described an incident that had occurred in our teaching (see below, the section on perspective-taking). The other participants in the meeting pointed out to us quite emphatically that our work was a form of evaluation-in-the-classroom!

By the time the discussion of these diagrams was over, the students had moved into round-robin phase: Every individual had done some work alone. Every student had given some description of his or her thinking. A general discussion had followed.

If we were doing this exercise again, there are some improvements I would like to introduce. First, there should be a more careful discussion of representation, i.e. how to make a good diagram. We might borrow Howard Nemerov's phrase, "image and caption." The students should come to see that their ideas can be translated into pictures or diagrams, and the pictures can be translated into captions. It would not be necessary to "instruct" the students to do this — the necessity for the caption, or more generally, for a multi-modal way of thinking and expression, could easily arise out of the class process itself.

Second, the structural diagram of the communication process is incomplete without another representation — of the evaluation feedback loop. Of this more later.

Third, it would probably be a good idea to introduce the distinction between real descriptions and idealized categories. This could have been done if we had forseen the variety of descriptions the students would give, and the consequent need to code and tabulate them in order to make sense of the ensemble of responses given by the class. In fact, we had underestimated the potential richness of the students' responses.

Fourth, it would have been good to couple the class exercise with some follow-up reading about research on communication patterns.

In spite of these regrets, it should be stressed that this class period worked well. The students got the idea of the course, and joined in the spirit of it. They got to carry out an exercise in representation, and to reflect upon it. And the teachers, for the nth time, were properly chastened by the once-again unexpected complexity and interest of the students' thinking.

At a workshop with college teachers (Douglass College, Rutgers University), we repeated the task of drawing diagrams of the typical college classroom. For the most part, the representations were similar to those produced by the class in Newark, although mainly of the B-type (some side-chains). There was also more explicit recognition of the presence in the classroom of some students who are "out of it."

MEMORY

It is hard to imagine high level productive thinking occurring without the involvement of the memory system. A thought must be held in mind long enough to be worked on and this process involves memory. Additionally, the productive thought must be retained long enough to be translated into some symbol system such as language or mathematics in order for it to be preserved and recognized as a productive thought. The symbol systems themselves involve tremendous loads on memory. Stated quite plainly, memory is a requisite of productive cognitive functioning.

Many interesting and important questions can be asked with respect to memory. Investigators have explored the areas of memory capacity, the nature of the memory trace, the structure of memory and numerous other aspects of the topic. However, in keeping with the course's major goal of helping people to think better, our classroom exercises have stressed the importance of the particular strategy used to store the "raw material" which is to be remembered. The choice of strategy can be shown to greatly affect the nature and amount of material that will be recalled.

People spontaneously generate different strategies for remembering material and one of the most commonly chosen of these strategies is simple repetition. For instance, it is possible to learn and remember the colors of the spectrum by repeating over and over again the color names red, orange, yellow, green, blue, indigo, and violet. The trouble with using this particular method of storing information is captured in the word "memorization". For many of us, the word "memorization" evokes recollections of intense boredom, feelings of resentment, and images of hackneyed poetry recited with military precision. Even if the lack of interest and enthusiam for the task can be overcome, the information acquired through repetition is generally isolated from larger, more coherent organizations of information. The critical question then is, "how can the individual acquire new information without depending on rote memorization?"

Various techniques for organizing unwieldy or unrelated material exist which can be used to facilitate the retention of unfamiliar items. Although these techniques are grouped under the term "mnemonics", the particular mechanisms of these various memory devices are quite diverse. With respect to the spectrum example described above, a good mnemonic for remembering the appropriate colors is to take the first letter from each color name and construct the name "ROY G BIV". This "first letter" technique, which is a good way of abbreviating information, is quite common and effective within a restricted domain. But therein lies the problem with all mnemonics. The arbitrary, artificial connections that are made between the items to be remembered are inappropriate for larger organizations of knowledge. The richness, complexity, and subtlety of such systems as Piagetian psychology, quantum theory, or the "self", cannot be reduced to a number of artificial relationships.

Despite the limitations of mnemonic techniques, they do accentuate an important point that the "raw material" of experience need not be "taken in" as it is presented, but can be worked on, transformed, and manipulated in various ways and to different degree. While the artificial, limiting aspects of mnemonics persuaded us not to pursue this topic through class exercises, we do emphasize the functional value of restructuring material; however, more meaningful forms of organization than those provided by mnemonics are explored. Each form of organization or "strategy" has a different impact on retention and each individual can bring his unique knowledge and experience to these organization tasks.

In pursuing the importance of what the subject does to the "raw material" of experience in order to remember it, we exploited the psychological experiment as a tool for illuminating this issue. After all, an experiment is designed to inform us about something and it therefore coincides quite well with pedagogical goals. The particular design we employed had three distinct conditions, and the students were equally divided among these conditions. All three groups were told that they should try to remember as many words as they could from a list that they were to receive. Group 1, the "Uninstructed" group was given a list of thirty-one words with no further instructions than those already given. Group 2, the "Instructed" group received the identical list but with the further instruction to organize the list in conjunction with trying to memorize it. Group 3, the "Pre-categorized" group received a list that contained the same words as groups one and two received, but the words were now arranged hierarchically. The superordinate category was "things" with the subordinate categories "alive" was further subdivided into the classifications "animals" or "fruit" while "manufactured things" were grouped as either "furniture" or "weapons". Several minutes were given for the students to work on their respective tasks.

The results of this demonstration are summarized in Table 2. The experiment proved to be quite useful in demonstrating the powerful effect of organization on recall.

Table 2

THE EFFECT OF ORGANIZATION ON MEMORY FOR A WORD LIST

GROUP	ITEMS RECALLED
Uninstructed	19.9
Instructed	22.3
Pre-categorized	29.4

In the next class session, we shifted focus to a slightly different aspect of the topic. Memory does not contain only those things we have explicitly tried to or have been told to remember. A myriad of facts and experiences reside in our memory system. How then do we account for which material is retained in memory? One factor that can be shown to have a strong impact on what we remember is the way in which we assign meaning to a particular action or task.

> Meaning is catalyst for organization. When we say that something is "meaningful", we are stating that it engages the well orchestrated system of interests, beliefs, prejudices, needs, etc. that form the organization we call the "self". Any experience that taps into this system will be organized in a more powerful way than experiences that remain isolated or independent from this organization. One would also expect the superior organization of "meaningful" material to result in enhanced recall for such material. In order to further explore the relationship between meaning and memory we again used an exercise

modeled after the standard psychological experiment.

A word list consisting of 45 adjectives was distributed to the class. The class was then divided into two groups by giving half the class one set of instructions and the other half an alternate set of instructions. The "Counting" group received instructions to "Look at each word, count the number of vowels in the word, and write this number next to the word." We expected this task to be regarded quite neutrally by the group.

In contrast to the "Counting" group, the "Self-reference" group received instructions to "Look at each word and ask yourslf whether or not it describes you. If it does, put a check next to it." We expected this task to engage the interest of the group.

It was necessary to prevent the students from intentionally memorizing the word list in order for the demonstration to be valid. With this in mind, the class was told that the exercise was designed to demonstrate a certain feature of language. Also, after the students completed the task, their papers were collected and five minutes of unrelated activity ensued.

Following this period, the students were requested to write down as many words as they could recall from the adjective list. When they completed this task, we asked each student, one at a time, to tell us the number of words he was able to recall.

There was a clear and dramatic difference in recall between the two groups. While the "Counting" group recalled an average of nine words each, the "Self-reference" group recalled an average of seventeen words each. Up to this point the students were not aware that there had been two sets of instructions and there was general puzzlement as to why there had been such a wide discrepancy in performance between the two groups. When both sets of instructions were made known to the entire class, there was a strong reaction on the part of many students. It became immediately clear that despite the fact that the "raw material" for each group was identical, the group that had examined the word list in the context, "does this word describe me?" had undergone a more interesting, personal, and meaningful experience than the "count a vowel" group. In this light, the discrepancy in recall performance between the two groups appeared reasonable.

An experience of this type usually activates the class and provides a good amount of material for discussion. The inclination, prompted by time constraints and the desire to arrive at a general synthesis, is to follow these memory exercises with a teacher led discussion of the issues. This procedure is exactly the one we employed. However, by following this conventional model, we probably short circuited the individual's process of discovery in developing his own strategies and techniques for dealing with the material presented in the classroom exercises. In retrospect, we should have utilized the round robin format to explore the diverse elements of the various constructions of the class members.

Both the demonstration concerning the effect of organization on memory and the exercise involving the role of meaning in memory clearly illustrate the essential point that the "raw material" of experience need not be passively registered. It can be transformed, manipulated, and digested. Strategies ranging from simple repetition of the given material to use of mnemonics, organization, and self reference represent different ways and degrees of making the "raw material" of experience your "own". The experiments we utilized give the student an opportunity to "try on" and evaluate the effectiveness of a number of these strategies through the immediate feedback provided by their own recall performance in relation to the recall performance of others using different strategies. It also provides the student with another opportunity to rediscover the fact that a vigorous, non passive orientation to cognitive life is important, not only with respect to memory but also in such diverse areas as problem solving, the recognition of propaganda devices, hypothesis formation, and productive thinking in general.

PROBLEM SOLVING

Within the pantheon of cognitive abilities is the skill referred to as "problem solving". Unfortunately the label "problem solving" can be deceptive. Problem solving is not a discrete, singular process that occurs the same way, every time, for everybody, for every kind of problem. This conceptualization of problem solving obscures the richness and subtlety of the process. Problem solving is more accurately conceived as a purposeful utilization of a variety of cognitive skills such as imagery, intuition, mathematico-logical thinking, etc., in a highly individualistic manner. Problems are also individuals. They vary in content, complexity, and in the time needed for solution. We chose to present problems that seemed capable of solution well within the time constraints of a single class session.

Although we explored problem solving through the use of fairly restricted problems presented one at a time in the hope that this simple situation would be conducive to an examination of the solution process, we will probably extend our focus in the future. It would be interesting to present problems that are of sufficient complexity to engage a student for a week, month, or even an entire semester. A task such as this would certainly better approximate how problems usually occur in real life. This does not mean that we should abandon the "half hour" problem, but that we should supplement it with problems of another scale.

Among the obvious forms of feedback in a problem solving situation is the actual solution or the response "right" or "wrong" from some arbitrary source. However, we shifted attention to an examination of the solution process itself. The class was divided into several small groups of four to five persons each, and the general instruction was to freely exchange their ideas on the problem and to keep track of how their thinking changed over the course of the problem, thereby constructing a reflective record of successive approximations to a solution. One of the problems we presented them with was a problem described by the psychologist Karl Duncker over thirty-five years ago. The problem is a follows:

> A person has a stomach tumor which cannot be treated surgically. A beam of radiation can destroy the tumor, but the beam also has the property of destroying the healthy tissue that lies between the beam and the tumor. How can this problem be solved?

Based on the responses from the class, the solution process seemed to fall into several discernable stages. At first, there were several requests for restatement of the problem in order to insure that they had "gotten it right". Following this "confirmation" phase, there was a period in which the majority of solutions either ignored or violated certain premises of the problem. For instance, replies were along the lines of "make an incision and focus the beam directly on the tumor" or "treat the tumor with chemicals instead of radiation". It was pointed out that while these solutions may be viable, they do not adhere to the limitations imposed by the problem. The problem explicitly prohibits surgery and implicitly excluded going beyond the historical or "state of the art" constraints, thus eliminating the chemotherapy option.

The next apparent phase displayed a strong tendency to concentrate on protecting the healthy tissue of the body, such as applying a screening salve to the skin or by shielding the body with a lead screen type of device. These solutions are not held for long because it becomes readily apparent that although they are successful in protecting healthy tissue, they correspondingly eliminate the ability of the radiation to effect the diseased tissue. Even if it were possible to allow the radiation to pass through the skin without harming it (selective protection), the problem of protecting the intervening internal organs would still remain.

At this point in the solution process an interesting thing occurs. Having had a "first go around" with the problem and coming up short of an answer, some students seek to distance themselves from the problem by giving up on it, or by concluding that some "gimmick" must be involved. The latter reaction seems quite reasonable in the face of the common past experience of having heard similar types of problems which turned out to have punch-lines instead of genuine solutions.

For those students who continue to pursue the problem (even the small setup does not prevent certain students from "dropping out" of the exercise), a curious shift takes place. A good number of the solutions now offered involve putting the patient's body into motion. The question arises, "What would happen if you rotated the patient's body so that the same spot on the outside is not continuously contacted but the same spot on the inside is continuously focused upon?" This line of reasoning represents a functional solution to the problem under certain assumptions. For instance, in order for this solution to be genuine, it must be assumed that the beam is weak enough not to cause damage under conditions of brief exposure (as is the case with the surrounding tissue) but strong enough to have an effect with longer exposure times (as is the case at the point of the tumor). Since we are interested in the solution process itself rather than getting the "right answer" we encouraged the class to continue with the problem. We informed the students that there was another, perhaps more elegant solution to the problem and that they should try to formulate it.

While several students reverted to earlier type solution in a slightly different form at this point in the exercise (i.e. put a tube down the throat and "pour" the radiation into the stomach), other students stayed with the notion of keeping the problem in motion. The critical development that occurred at this time was a shift in attention from the body to the radiation.

The first solution offered after the shift of focus to the radiation is the converse of the "rotating body" solution. This new solution involves holding the body in a constant position while the beam is rotated around the body. Although this solution is very close in form to the "rotating body" solution, the ground-work has been set for a "final" explanation. The problem has been firmly established as one of focusing a beam on an inner location while protecting the surrounding regions. By putting the problem into motion, the critical idea of changing the location of where the beam contacts the body has been brought into play. Attention has also shifted to the beam itself. The realization soon comes that there is another way of changing the location of the beam. This change involves not a successive change in location, but a simultaneous location change through the use of multiple beams at the same time. The idea of lowering the intensity of the individual beams in order to meet the requirement of effectively treating the tumor while protecting the surrounding tissue follows fast on the heels of the multiple beam notion. The problem has been solved, but more importantly a unique opportunity to critically examine a "piece" of thinking has been provided.

TAKING ANOTHER'S POINT OF VIEW

The act of seeing things from another person's point of view is a central theme of the whole course. In almost every class meeting there is an opportunity to do this and to reflect on the results. But we wanted also to do some work more directly aimed at becoming aware of the process of perspective taking. In the fall of 1977 Camille Burns and Howard Gruber planned a three-unit sequence with this end in view. The plan was as follows:

a. Understanding poems in which the meaning turns on a sudden shift in perspective. We planned to have the students read first a very simple poem and then a more complex one. After they had understood each, the next task would be to discover what they had in common (i.e., sudden shift in perspective).

b. Struggling with moral dilemmas in which the question of what is right depends on whose ox is gored. The moral dilemmas were brief anecdotes of the kind invented by Piaget and by Kohlberg to study the development of children's moral judgment.

c. Writing a dialogue about a perplexing social issue in which the student is required to shift perspectives as he or she first writes one speaker's lines then the other's.

It should be stressed that we were distinctly not trying to inculcate a 1950s social-science "objectivity" or non-partisanship. On the contrary, when the time come, we tried to bring out the idea that understanding other people is important in order to struggle well for what you believe: to clarify your own ideas, to discover your allies, to anticipate your opponents. But the first step in all this is to understand the other.

Complex plans are risky in a teaching process predicated on inviting the students to think. What if things don't go as expected? Must the whole plan go out the window? In this instance, that was almost what happened.

The first poem we used was "Quatrain" by Sarah N. Cleghorn:

The golf links lie so near the mill
  That almost every day
The laboring children can look out
  And see the men at play.

We expected it to be very easy to understand. Half an hour at most of

Quatrain and we could get on to Shelley's Ozymandias. That should leave plenty of time to finish the period by collectively discovering what the two poems had in common.

To my surprise, Quatrain turned out to be very difficult to understand. The class was unfamiliar with the term golf "links" -- that was easy: I translated it as "course". They didn't quite grasp that golf was a rich man's game -- maybe it isn't so much any more. They didn't know about the history of the struggle for laws prohibiting child labor. And probably, they didn't have a clear idea that the function of poetry might be to voice social protest. These points emerged as we went around the room and each student gave his or her interpretation of the poem. As each student spoke, betraying a wealth of unlooked for misunderstandings, my dismay grew. I was dejected, not so much at the plan going wrong, but at the low level of culture I perceived in the group. And I was on the verge of committing what would be the cardinal sin within this method of teaching — simply lapsing into telling the class "right answer".

But I persevered. I provided some cultural and historical background; the class categorized their different interpretations, and then discussed them. My interpretation was eventually included in the list, but I did everything I could to avoid suggesting that a poem has only one right meaning. The period ended on a note of irresolution.

The next period, I was still tempted to go back to Quatrain and insist on the right answer. I resisted the urge -- we had had enough of those four lines for awhile -- and we went on to Shelley's Ozymandias. To me, this poem seems more difficult than the Quatrain: longer, more complex, more exotic. But on the whole, the class understood it quite well; that is, their understanding matched mine fairly closely. I had learned my lesson and we took our time. Listening to the nuances of their differing reactions I heard things I had never noticed in the poem, although I have known and re-read it over a span of 40 years. The students moved me by their insight, and my spirits lifted. But we ran out of time and still had not gotten to the question originally planned: what do the poems have in common? (Right answer: a sudden shift in point of view.) I asked them to write out a paragraph for the next class, dealing with this question (they had copies of the poems).

At the third period in the sequence, we went round the room again. Yes, some of them got the "right answer". But far more important, some of the students discovered something else the poems have in common: they both deal with power: So by the time we were done with this, the students had thought about two poems, perhaps more carefully than ever before in their lives. They had seen the many interpretations possible, both widely ranging and sounding many fine nuances, making the ideas of the person next to you worth hearing. I had learned a lot from the group, and my interest in their ideas was important to them. And, albeit a bit slowly, we had come out of it in a reasonably good position to go on with the original plan. The next steps went very well. The second task, moral dilemmas, were marvellous grist for the mill of our circular process. The third task, dialogue-writing, was difficult but not overwhelmingly so. The class was mostly Black and Hispanic. I had chosen as the material for the dialogue a letter that had appeared in the student newspaper, evidently by a white person, arguing against affirmative action programs in which minority group members are given preferential treatment in employment practices. Everyone found it easy to answer the author. But the task we had set was to answer the author, then to give the author's reply, and finally to have the last word. Some of the students were reluctant or found it difficult to frame a real argument for the opponent's side, and resorted to having the opponent say merely, "You're wrong." In comparing the widely varying productions of the students, the weakness of this strategem became evident without my pointing it out. There was a difficulty, then, but the class overcame it.

That year, the students' interpretations of the poems were given orally and I have no exact record of what they said. In 1979 Bob Keegan and I used Sarah Cleghorn's Quatrain as part of a somewhat different exerecise which began with having the students write out their ideas of what the poem means.

This group came a good deal closer to agreeing on the interpretation of the poem as a protest against child labor, possibly because of the way we set the stage for the exercise. Nevertheless, the interpretations cover quite a range.

STUDENT INTERPRETATIONS OF SARAH CLEGHORN'S QUATRAIN

M.W.

Children are not the only ones who play, or fool around, but grown-ups also have the need to enjoy and play at some point or other.

N.H.

The poem suggests that children are working but not so far off they can see men playing. This could mean that young people are going through certain cycles so that they can be where these men are. For instance going to school to get an education will soon earn young people jobs that are being held by adults at the present time.

J.W.

Children are hard at work, while grown men can find the time to play golf. This is as if the men want to rub the children's faces in their poverty.

R.L.

It seems as if the golf course is so near the mills, or working area, that the children who are working can see the men playing golf.

V.L.

The point of view is reversed. The men should be working (laboring) rather than the children. The children should be playing rather than laboring and watching the men play on the golf links (golf course).

L.W.

(Summarizes poem about same as R.L.): I think it would be better if the men were working and the kids were playing. Or at least the mill shouldn't be so near to a recreation center, because it would make the kids feel sad.

J.B.

From children's point of view, the poem seems to suggest that looking out to see the man at play is something that is taken for granted. The children can see their view which includes the golfers' view too. View not absolute as it encompasses two views too. It is a relative view.

R.S.

During their chores the children observe the older men playing golf. Makes me think of inequality and bondage.

L.M.

(edited slightly) I gather, from the word mill that they mean something to do with wheat or grains, because that's what goes on at a mill. It doesn't take much of anything to work at a plain old mill. From laboring children I get the impression that the author is talking about slaves. The men at play are rich men (white), play golf watching the poor children (slaves — black) working. I mentioned earlier that it doesn't take much of anything to work in a mill, meaning the stereotype that blacks only had muscle and no brain so only labor jobs would be issued to blacks (dummies). And realistically the game of golf is played mostly by white people.

B.B.

There are wealthy men playing golf on a golf course. Near the course was a mill where poor families sent their children to work at the mill to help support the family. And everyday the children look out at these men and wish they could play and not work.

M.D.

Things are reversed. Instead of the men working and children at play, the children are working in the mill and the men are out playing on the golf course.

P.R.

I think that this poem is saying there is a certain irony between the children laboring and the men playing. Usually you would expect this to happen in reverse. This also suggests that the children are poor and the men are rich.

G.R.

This poem was written many years ago. Children didn't go to school because they had to help support their families and themselves. But the men who were well off could have leisure time to do things such as play golf.

B.S.

All the subjects were adults, but were categorized according to their wealth...Children as opposed to men symbolize the superiority of the playing group...

K.S.

...a group of children who are busy at work...can see men playing golf, near where they are laboring.

D.R.

It means that while the poor children are working every day, they probably wish they were playing. Instead, they just watch the men relaxing or playing, while they are working hard.

T.C.

Children in the days of the depression worked in sweatboxes. These children work in a mill. Looking out of the factory windows these childlren are working at normally men's jobs. The men are playing a childish game, in this case golf.

M.F.

Children are hard at work, while men are busy at play. Should be the other way around.

P.C.
...men playing golf on the golf links because of the labor of children.

EMOTION, COGNITION AND REALITY

Our basic strategy is to single out some aspect of cognitive functioning, develop a task situation that calls upon that aspect or sub-skill, and draw the students' attention to that domain. But any real performance, of course, draws on varied kinds of knowledge and skill. The focus of the course depends not only on the list of tasks proposed and on the unpredictable interplay among all those invlved, but also on the teacher's emphasis in steering the class one way or another.

While our main emphasis is on intellectual functioning itself, we are aware of the vital relation between cognition and emotion. This relation becomes paramount as one tries to think as well as possible in real, human situations. In different ways, some of the exercises we use are aimed at increasing awareness and control of this relationship. We give here only brief indications of some of our efforts in this direction.

ANGER. We ask the students, "try to remember some incident in which you were angry at a teacher." Sitting quietly, each student writes out notes on his or her recollections. Usually, no memories come at first. After a few minutes they begin to pour out. Then we go around the table with each student reporting. This turns into a very lively discussion and could occupy many weeks if we let it. As the session goes along we steer attention toward how the students had handled the situation in which they found themselves, and eventually to reflection on the availability of alternative strategies for coping.

BEWILDERMENT. This experience grew out of a planned activity that was side-tracked by the spontaneous course of events. It might not be repeatable, but the general idea is interesting. One semester, we wanted to draw the students' attention to how they listen to a lecture, take notes, and use those notes. We had a plan for this sequence which we never completed. The first step was for the students simply to obsereve themselves in any other class and to come to our class prepared to describe how they listen. When we had the round-robin, it became clear that they all felt bewildered, overwhelmed, baffled, and finally bored by most of their lectures. They felt the teachers were "snowing" them, and not paying attention to the students' needs.

We offered them a choice. Either we could try to work out ways to listen as well as possible in such situations - in our view, not an entirely unrealistic plan, since so much of life is like that. Or we could work out ways to try to change the situation. The students chose the latter path. Together we worked out a simple plan - nothing more than raising the issue with the teacher in question, either before, during, or after a class period. Each of our students took on the responsibility of trying to change a class!

This was one of the few times that most of the students in our course failed to do their homework. A few did do it, and everyone's reflections on the difficulties experienced in doing or not doing it were of great interest.

INTRODUCTION TO PART IV

In this section we turn our attention to a reasonably concrete illustration of assessment that departs from present practice and is consonant with the views expressed earlier in this volume.

It is by now clear that the members of our study panel hold the view that the distinction between assessment and instruction is largely artificial and arbitrary. Hilliard (1980 personal communication) for example in discussing this point says:

> "A testing and assessment system can be built without direct reference to learners. When this happens, the "correct" logic and content of answers to questions are assumed to be known in advance by the questioner. The goal of testing in this case is to determine if learners agree with questioners. A testing and assessment system may also be built to use the learner's repetoire for building questions. This has sometimes been referred to as response-contingent testing. THE KEY POINT TO BE MADE HERE, HOWEVER, IS NOT A POINT ABOUT TESTING PER SE. IT IS REALLY A POINT ABOUT TEACHING. Any type of testing which is selected will fit a particular philosophy of and approach to teaching. Paolo Freire has described two different approaches to teaching. The "banking" approach is generally manipulative. Student are said to "learn" when their answers to questions match those with which the teacher begins. An alternative approach is called a "dialogical" approach. Students are said to "learn" under this approach when they become problem-posing activists. Both questions and answers are new to both teachers and students. In the first approach, the teacher's role is to "donate" the material which the student is to learn. In the second approach, the teacher's role is to establish a true dialog between teacher and student.
>
> These are not mere theoretical matters. Paolo Freire is astoundingly successful in using dialog to teach literacy and problem solving. William Johntz and teachers who are trained by him are equally successful in teaching low income children, from any cultural group, relatively abstract mathematical concepts and skills where others had failed to teach arithmetic before. In both cases, "testing" or assessment is ongoing. The teachers and students use the students' repetoire as the building blocks for learning...
>
> Here are some examples of ongoing assessment... Paolo Friere places great stock in listening to his students. He listens to detect those things about which they have strong feelings. He listens to record the vocabulary which his students already know. These two parts of his systematic assessment process are then used directly in instruction. Students learn to read (in about 30 hours of instructional time) by using their own words and by focusing on issues of importance to them. William Johntz and Project SEED teachers place great stock in listening for student logic and for student assumptions. They also listen for full participation of all students. They observe exactly where each student agrees or disagrees with each step of the group's problem solving effort. They observe if students are willing to argue for positions which they hold, even if alone against the entire class. These and other data are collected

systematically in order to design the "in flight corrections" of the teaching strategy. William Johntz and Project SEED also direct a part of their ongoing systematic evaluation of the instructional process toward the teacher. Peer critique is always used, and it is done systematically.

The intertwined nature of assessment and instruction requires that neither be secret. Secrecy is anathema to education. If assessment is to indeed serve instruction, then ways must be found for students, teachers, and parents to derive insight and information from the assessment practices that are in fact employed.

In the context of assessment in the service of instruction a key argument in favor of open assessment practice is best put forward by Taylor. (Times Educational Supplement, London, Nov. 16, 1979) who says;

> We all have important stakes in the results of the tests our society administers. Some of us are directly involved: students, parents, teachers, school administrators, testmakers, and the public. Other interested parties look on with more or less political power: public administrators, legislators, academics, critics, columnists, and interested citizens and taxpayers.
>
> Secrecy of tests erects unnecessary walls that hinder the many-sided interchange among all these interested parties. Secrecy aggravates inequalities that already exist, for instance between administrator and student, or between testmaker and critic.
>
> In our society test results are taken to be indicators of success and worth for individuals and school systems. That is what makes the secrecy of these tests so uniquely perverse and damaging. Since that secrecy is also unnecessary, its elimination should have a high priority in public discussion and public policy.
>
> Freedom of information acts and other legislative remedies are steps forward, although few parents, for example, have the knowledge, determination, or resources to invoke such laws. It may be that test secrecy will finally be eliminated only after major court cases result in substantial damages being paid to some of those who now suffer their injuries in silence.

In addition to openness, it is clear tha new approaches to assessment in the service of instruction demand a re-examination of the notion of "validity" as applied to the design and use of assessment instruments. In the paper by Schwartz, Taylor and Willie, the work of project TORQUE on this important methodological question is presented.

Project TORQUE was a research and development effort supported by the Carnegies Corporation and The Ford Foundation that was charged with the responsibility of designing alternative assessment techniques and instruments for elementary school mathematics. In the course of this six year project, a different approach to validation was evolved; one that was not correlational or even statistical, but rather categorical in nature. Such an approach to validation seems to have produced techniques and instruments largely free of the flaws of more traditional approaches.

The paper describing project TORQUE that is presented in this section clearly does not constitute a handbook to the educator who is concerned with the need to devise assessment techniques that are informative, non-threatening and open. We hope, however, that it will provoke careful consideration of what we believe to be some important principles to be considered in any such effort and that explicitly guided project TORQUE in its work. They are:

> In an instructional context, distinctions between assessment and instruction are arbitrary and artificial.
>
> Valid inferences about a person's knowledge can only be made within a framework that incorporates an understanding of the task and the students idiosyncratic representation of its structure.
>
> Students' introspection, reflection, self- examination and seeking out of critical appraisal are all evidence of successful educational experiences.
>
> Intellectual progress, whether of an individual or of the species, is impeded by secrecy.

Project TORQUE

An Example of Categorical Test Validation

Judah L. Schwartz
Edwin F. Taylor
Nancy A. Willie

I. INTRODUCTION

People who make decisions about other people's lives have social and political power. Insofar as testing is used to influence these decisions, tests are instruments of power. The pervasive use of tests in the United States has bred much criticism (Houts, 1977; Hoffman, 1967). This criticism has had some results: advocacy groups, educational reform movements, legislation, and regulation, all of which seek, by one means or another, to protect the rights of individuals from "the tyranny of testing."

In the short term, scrupulously responsible use of currently available tests may help meet such criticism, but a long-term solution requires more fundamental reform of test development and use, a reform whose seeds may now find fertile soil.

The work of Project TORQUE* described in this paper was motivated and guided by our concerns about the role of testing within the larger societal contexts in which it occurs. A pluralistic and democratic society requires tests that are subjected to the scrutiny of many "experts" and the public-at-large, for whom testing has social, political, economic, educational, and ethical consequences. We write for those who share our interest in education and society, and not only for professionals in the fields of education and testing.

This paper outlines the foundations and traces the consequences of several assertions:

(1) Testing cannot be separated from an understanding of the task being tested. Test-making is, in large part, the act of seeking understanding of the domain being tested and of the ways people demonstrate their leaning of that domain.

(2) Some learning domains can be analyzed into tasks and subtasks on which performance can be observed and categorized as "all-or-none."

(3) "All-or-none" subtasks, when they arise empirically rather than arbitrarily, are useful in describing performance (testing) and helpful in improving performance (instruction).

We apply our theory of performance categorization to the domain of mathematics learning, specifically to the tasks of making measurements of length, area, volume, weight, and time, and to the development and validation of several tests of performance on these tasks. (N.B. We reserve the word MEASUREMENT for the application of numerical scales to physical quantities. In evaluating human performance, we try to use categories rather than numbers.)

---

* A research group at the Education Development Center, Newton, MA 02160, supported by the Carnegie Corporation of New York and The Ford Foundation.

The following sections outline our theory about "all-or-none" performance categories, discuss the implications of that theory for test development and validation, provide an account of the process we designed for test development and validation, and consider the generalizability of our work.

One characteristic common to most testing practice is the reporting of a test result as a number on a scale or a score. We believe that this application of numbers to an individual's skills and performance is unjustified, and that the use of numbers in this way confounds and misdirects educational endeavors and the development and use of tests. We outline briefly the arguments in order to set a theoretical stage for the categorization of performance described in the following sections of this paper.

In the natural sciences numbers are used to describe two kinds of quantities. Discrete quantities, such as the number of apples in a basket or people in a room, are countable. Continuous quantities, such a the distance from Boston to San Francisco, are measurable. The followng acts are necessary elements of any such measurements;

> Identifying the attribute of the object to be measured, and distinguishing it from other attributes the object may possess;
>
> Choosing a unit of appropriate attribute and size;
>
> Comparing the attribute to be measured with the unit;
>
> Judging a level of precision appropriate to the context of the measurement.

We do not consider any situation in which the attribute is defined only in terms of the instrument used to "measure" it as being an example of measurement. Thus, Boring's "IQ is what IQ tests measure" is in our view, at best, tautologous. The attribute to be quantified must have some independent definition.

Assume for the moment that it is possible to identify a distinguishable attribute possessed by an individual and that one wishes to measure it. Is it possible to define a "scale" that can be applied to the attribute? The use of numbers to describe quantity rests on the assumed existence and appropriateness of such scales.

Traditionally, scholars have referred to nominal, ordinal, interval and ratio scales as being suitable for the measurement of psychometric variables. We believe that only the ratio scale is a scale that permits measurement. Neither nominal nor ordinal scales have anything to do with measurement except in a loose metaphorical fashion. Nominal scales simply assign numerals to objects on the basis of whether or not the object possesses a paticular attribute. For example, a nominal scale could assign the number 7 to all objects that are pink and the number 10 to all objects that are green. Ordinal scales simply rank-order objects according to the amount of an attribute which they possess. For example, glass can scratch steel, and steel can scratch wood. Thus glass, steel and wood might be assigned the numbers 1, 2 and 3 respectively, because they can be ordered by "hardness". The interval scale concerns itself with the intervals between the extent to which objects possess an attribute. Standard intervals, called measurement units, are defined in terms of the standard of comparison. A common example of an interval scale is the Centigrade scale for measuring tenoerature, in which the difference between 10 degrees and 20 degrees is equal to the attribute being measured. The arbitrary zero point of the Centigrade scale should not be confused with the fact that there does exist an absolute zero the temperature scale i. e. 0 degrees Kelvin. The existence of a non-arbitrary zero point which implies the ability to distinguish in a categorical fashion the presence of the attribute frm its absence, is, in our view, central to the identification of the attribute. Only a ratio scale has this characteristic.

A ratio scale, has the following properties. First, there is a non-arbitrary zero point. Second, the ratio scale can only be applied to one dimensional attributes. One cannot order unambiguously points in spaces of more dimensions. Third, one must be able to quantitively define what is meant by the interval, a "little bit more", of the attribute. Units, such as "one degree hotter", or "one centimeter longer", or "one second later" must exist. Ordering is insufficient; scaled comparison is necessary. Without such scaled comparison, measurement can have no consistent numerical outcome.

The concept of "a little bit more" cannot be quantified and applied to individual human performance, even in cases when highly refined and specific subskills are identified as the attribute. For example, we identified the "subtask" of usng the zero point on a ruler correctly when measuring the lengths of lines. For this subtask, performance can be described by performance fractions, (the number of correct uses of the zero point)/(the number of opportunities). Performance can be ordered: 4/10, 5/10, 9/10 and 10/10. One must be able to say how the interval, say, frm 9/10 to 10/10 compares in size to the intereval from 4/10 to 5/10. The intervals themselves must be capable of being ordered if there is a true ratio scale. Is the student who gets 5/10 correct superior in the subskill to the student who gets 4/10 correct by "an equal amount of superiority" as the student who gets 10/10 correct is to the student who gets 9/10 correct? Degrees of superiority of human performance have no unique meaning. Without this unique meaning, all scaled performance, whether in comparison with other test-takers, or in comparison with a "perfect" performance, is not appropriately described by a ratio scale. And, thus, it is not capable of being measured.

II CATEGORIZATION OF PERFORMANCE

In our empirical investigations of the tasks of measuring extensive physical magnitudes, we have found subtasks on which people's performance is consistently either present or absent and which permit us to replace metric measures of performance with categorization. This section describes how we analyze tasks into such "all-or-nothing" subtasks and what happens when we cannot do so.

Our model of measurement, which derives from the physical siences, identifies the following major steps in making measuremnts:

(1) Identifying the attribute of the object to be quantified,

(2) Choosing a unit of appropriate size,

(3) Carrying out the comparison of the object to the chosen unit,

(4) Judging a level of precision appropriate to the context in which the measurement is made,

(5) Reporting the results.

We identified subtasks for measuring length, time, area, volume, and weight during an iterated process of theory formation, task analysis, and empirical trials with students and teachers in elementary schools. We used the five-step model of physical measurement to inform an initially rather unfocussed exploration of a given task such as length measurement, with students and their teachers until we began to notice parts of the task on which students performed either well or not at all. These "parts" or subtasks were progressively refined and gradually embodied in games and activities and some test "items" that allowed the beginning of ordered performance data. The measurement model was continually invoked and refined to help us decide whether or not our set of subtasks was relevant and comprehensive.

A sufficient task analysis would yield ordered performance data for each subtask. These data would cluster in "consistently present" and "consistently absent" categories, with few in the inconsistent category. We took the existence of this dichotomous categorization to be evidence of distinguishability of the given subtask.

The observed dichotomous categorization allowed us to dispense with the scoring of performance: everyone's (or almost everyone's) performance could be categorized within the present or absent category. Thus ordered performance collapsed into two-valued categorization.

In summary, the process of arriving at a categorized description of human performance included developing an understanding and model of the task, increasingly focussed activities with students and their teachers, verification of dichotomous performance on subtasks, and categorization of this dichotomous performance; the entire process repeated cyclically until successful.

Or unsuccessful. For some skills we were unable to identifysubtasks that gave rise to "all-or-nothing" performance. In particular, the task of computing elapsed time intervals frustrated our attempts at analysis and categorization. It may be that we have not been sufficiently insightful or persistent. Or it may be that for some tasks the subtasks are so interrelated that performance on one subtask influences performance on another. Or finally, of course, this result may indicate a failure or region of inapplicability of our method.

One final remark is in order before closing this section and moving on. It is not possible to completely separate or unconfound the effects of the observer and the phenomenon being observed. We know this to be true in the physical sciences where the assumption that experiments may be repeated and that the nature of the interaction between the observer and the system being observed is known and is small are plausible. In the course of observing human intellectual behavior it seems to us that these assumptions are rather more questionable.

Methodologists have written extensively on this subject (see for example Campbell and Fiske ( )) attempting to resolve the issue. We have tried to follow the spirit and intent of their procedures but in the end we hold the question to be non-resolvable, i.e. there can be no complete unconfoundng of "method and "trait". We present our results along with our methods as objectively as we can, and hope the reader draws a similar pattern of inferences from them.

III TEST DEVELOPMENT AND USE

Our use of performance categories instead of scores led us to modify traditional notions of test validity. We consider here some topics that affect the meaning of "validity."

TESTS VS. "REALITY" OR WHAT IS SOMETIMES CALLED "CONTENT VALIDITY."

Tests, as close observation for some purpose of an individual's performance, can consist of actual performance on a task, such as swimming or driving or doing arithmetic calculations, or can consist of simulations or representations of "actual" tasks. Such representations are useful because it is not always practical or possible to observe and test an individual's actual performance in natural settings.

When tests are constructed to represent "reality," the adequacy of the representation is of critical concern. Ultimately, there is no way to prove that a test examines performance on those tasks that it claims to examine, because there is no way to be sure that validating tools are themselves "valid." However, we believe that when tests are developed in the settings in which they will be used, when such development is the result of extensive observation of "real" performance as interpreted using a model of that performance, when tests closely resemble performance on alternate simulations of "reality", then one can feel some confidence that the test examines the skills one wishes to observe. We developed games and activities to stand in for "reality" during our validation process: these games and activities permitted us to observe the consistent presence or absence of performance on the subtasks of a measurement task in several settings and to provide a context in which motivation was reasonably high.

CONSTRUCTED VS. SELECTED RESPONSES OR WHAT MAY BE CALLED "RESPONSE VALIDITY."

Just as one must be concerned about the adequacy of a test as a representation of "real" tasks, one must also feel confident that the ways in which people respond to test questions represent the ways that the undertake the "real" task. Our concern with the responses of students had led us to reject the selected response ("multiple choice") format for several reasons. First, constructed responses simulate real performance more realisticaly than selected responses; people do not ordinarily choose among possible answers when measuring length or time. Second, constructed responses allow students latitude in the ways they can perform tasks, a latitude especially important in a pluralistic society. Third, a constructed response can signal the presnce or absence of performance on several subtasks, which often can be separated using evidence from the detailed response. Fourth, constructed responses permit a diversity of errors, from which teachers can refine their understanding of performance and non-performance in order to make instructional decisions. In addition, constructed responses provide us with a stringent check of our understanding of the task being tested and of the presence of task-analytic categories. The likelihood of students performing in predictable ways is greatly reduced when constructed answers are permitted. When our model of the task does account for the diversity of constructed answers, we can be more satisfied with our understanding of complex behavior.

THE USERS, OR WHAT MAY BE CALLED "PRACTICAL VALIDITY."

People use tests for many reasons. The tests described in this paper were designed to permit teachers (and others with similar concerns) to describe individuals' performance and errors in one domain, measuring. Students may learn to make sucessful measurements as a result of experiential learning that does not compartmentalize tasks and subtasks, but the teachers' role as "trouble-shooter" in this learning process requires that they have some analytical approach to the performance of their students. Teachers need to identify those students who show "mastery" of the skills of measuring, and of equal importance, they need to be able to characterize the needs of those students who have not demonstrated "mastery."

"All-or-none" performance makes some instructional decisions relatively easy. The measurement tests developed by Project TORQUE each prove the student's skill in only one kind of measurement. On each test, regardless of the number of "items" (between 6 and 12), the general criterion for mastery is one or no errors. A student who meets this general criterion has made, at most, _one_ error on _one_ subtask. For those students who do not meet this criterion, another look is necessary. This second look and the consequent categorization of the errors, is a rich source of useful diagnostic information. In some cases, the error analysis may reveal that the absence of performance on only one subtask is the source of several errors. In other cases, the error analysis may reveal that all the measuring subtasks have been mastered, but that errors have been made in related tasks such as counting or calculating. In still other cases, this second look may prove insufficient, and a third look with an alternate version of the test or with games and activities like those used during test validation, may be necessary before a teacher can decide on a student's learning needs. (To facilitate this process for teachers, we provide them with information we have gathered during our research and development work. A clear description of the subtasks and common errors of measuring is written in a teachers' manual. A list of the categories of common errors and a partial list of commonly occurring wrong answers which signal those errors, is printed directly onto a teacher's carbon copy of each student's test.)

The descriptive power of tests which are based on "all-or-none" categories may have significant instructional results. In preliminary field trials of the TORQUE measurement tests, teachers have been able to identify specific learning needs and to focus instruction on them because they have been able to observe their students' performance and to interpret that performance in terms of a theoretically derived and empirically verified model of the task.

IV DETAILS OF THE DEVELOPMENT PROCESS

This section gives a detailed account of our process of test deveopment, using as an example one test of the measurement of area.

After observing students, reviewing current curriculum materials and extensive discussions in our staff and with classroom teachers, we designed games and activities that permitted us to observe performance on the tasks of area measurement according to our general model of measurement. Students then used these games and activities in their classrooms. Although teachers and students were enthusiastic about the games, teacher observations did not identify subtasks on which dichotomous performance could be observed. Staff members then worked intensively with small groups of children, using a variety of trial materials, until we had focussed progressively on subtasks on which performance seemed to meet our criterion of dichotomy.

We found that for area the identification of the attribute, the first step in our model of making measurements, was a difficult task for many students, and that we could describe subtasks related to this step. As a result, we decided to design two tests of area measure: the first test focusses on attribute identification by asking students to measure areas by "covering" regions with a nonstandard "tile" unit; the second test deals with the measurement of area by computation using measured lengths. In this section, we trace the development of the first of these tests.

Our accumulated experience with teachers and students revealed two major subtasks of identifying the attribute of area:

> (1) DISTINGUISHING BETWEEN LENGTH AND AREA. The most common error students make is to use area units, in our case "tiles," as units of length rather than area. (We found this to be true even when area units were triangles or hexagons. The longest length of the area unit was used as a length unit.) When presented with rectangular and irregular regions, some students measure one length, others add two perpendicular distances, others measure the distance around the edge of the shape, the perimeter.
>
> (2) DISTINGUISHING BETWEEN SHAPE AND AREA. Many students do not distinguish between area and the shape in which it occurs. When we presented regions that could be "covered" only by using half-tiles as well as whole-tiles, some students ignored those parts of the region which could not be covered by whole-tiles, other students counted every half-tile as a whole unit, and still others used overlapping tiles, each counted separately, to avoid partial units altogether.

Performance on these two subtasks could be observed as students "covered" a variety of regions, some of which could be covered by whole-tiles and some of which required both whole-tiles and half-tiles. Length confusion was observable in both cases, while shape confusion was observable only in the latter. We labelled the subtasks "whole units" and "half units," for convenience. There are peripheral tasks, such as counting, adding, and familiarity with fractions, which we knew would affect performance, but we believed we could separate performance on these peripheral tasks from performance on the attribute subtasks.

Several preliminary versions of an "initial area test" and a wide variety of validating games and activities were piloted with small groups of children. We sought a set of materials that would demonstrate simultaneously (1) that our analysis of the subtasks was sufficiently correct and specific so that children performed either all-or-none on a given subtask, (2) that a significant fraction of the constructed incorrect answers occurred in anticipated patterns, and (3) that test performance on each subtask was consistent with performance on validating activities.

One validating activity which evolved from this procedure is shown in reduced form in Figure 1. It carried the English title, "Lots of Land," and the corresponding Smanish name, "Ranchos Anchos." Students considered it a map of plots of land which they could purchase by measuring the area of each plot in tile units. Starting with any plot of land along a short side, the player would move from one lot to an adjacent lot, measuring the area of each one, until a connected path crossed the board. Players were asked to pass through a "free" lot in the middle of the board in order to insure that every player would measure a sufficient number of each kind of area chosen according to the subtasks listed above.

Figure 3 shows one form of the test that was validated against the "Ranchos Anchos" activity. The apparent simplicity of this test is somewhat deceptive: every graphic feature and characteristic of each item is the result of much close observation and many discussions. Behind this particular version of the test is a set of rules for generating each item in multiple versions.

The six items of the Area Measurement Test (tile units) have the following characteristics:

> Items a and b are rectangles which can be covered by whole "tile" units and which contain interior cells.
>
> Items c and d are irregular, contain interior cells, and can be covered by whole "tile" units.
>
> Items g , h, i and j are irregular, contain no interior cells and must be covered in part by half-units. The regions in items (g) and (h) which must be covered with half-units are more easily partitioned than the half-unit regions in items (i) and (j). For items (g) and (h) the half-units can appear as tabs, with three aides exposed on the contour of the shape. For items (i) and (j) the half-units are embedded in the shape.

Two versions of the test were used in the validating procedure described here.

The validation process itself took place in various school systems in which we could visit classrooms with children from diverse ethnic, cultural, linguistic, and socioeconomic backgrounds. A typical validation session went as follows: two to four staff members would appear in a classroom at midmorning with a box of materials. Each staff member (rather than the teacher) would select two children at random and sit down with them at a table to one side of the on-going classroom activities. The staff member would explain to the children that we were making up tests and needed their help to discover whether the tests were good enough. We told them that we would be taking notes on observation sheets while they did the test and played some games, and that we might ask them some questions as they went along in order to understand their thinking. When the session was over, we would answer all their questions and talk with them about what we and they had learned.

The children would first take one version of the test as pretest. During this and the entire validation procedure the observer would watch the measurements being made, take notes, and ask for explanations of strategies that the children were using. After the test came the validation activity, in this case a one-person game, although for tests of some other measurement skills the games were group games. Following the game, each child took an alternate version of the test as posttest. (Our terms "pretest" and "posttest" refer to their position as first and last in the validation procedure. This use of the terms is not to be confused with conventional uses in which explicit instruction takes place between the two tests.)

After the formal validation procedure was completed, we welcomed the children's comments and criticisms. We refrained from making judgments about individual student performance, but we encouraged discussion about the questions and the activities. When children showed specific interest, we spent some time teaching them about the measurement skill that was the subject of the validation activity. We showed the teachers copies of the tests and validation activities but did not discuss with them the performance of individual children.

These vaidation sessions typically lasted about an hour each.

Now began the work of interpreting the observations. Each observation sheet, along with the completed written tests and game sheets, carried as full an account as we could manage of the behavior observed. We organized this account under the subtask to which we wished to pay particular attention. The pre-condition of validation was "all-or-nothing" performances by a large majority of children on each subtask in both tests and validation activity. The criterion of validation was the consistency of performance on tests with performance on the validation activity. In the following sections we report on the application of this precondition and criterion to a variety of tests.

V DETAILS OF THE VALIDATION PROCESS

Our analysis of each student's performance on each subtask began with our interpretation of each constructed response, interpretation made reliable by our observations and notes. We made a decision of "correct" or "incorrect" for each subtask included in each response. A "performance fraction" (number of correct responss/number of opportunities to respond) was used to describe the subtask pereformance for each student on pretests, validating activities, and posttests. Performance fractons were plotted on linear scales, as shown in Figure 4. When most performance fractions fell near the "top" or the "bottom" of the ordered scale, we had satisfied our precondition that correctly identified subtasks give rise to dichotomous performance. Table 1 shows, for each subtask on the series of measurement tests we designed, and for both methods (tests and validating activities), the percentage of students whose performance fractions on subtasks fell within the "top" and "bottom" boundaries, thus satisfying the precondition of dichotomous performance.

Defining the boundaries of "top" and "bottom" performance is central to our process of validation. By defining "top," "bottom," and "middle" regions on the linear scale, we were able to categorize all the performance we observed: consistently correct, consistently incorrect, and inconsistent. How high is "top" performance, and how low is "bottom" performance? We examined this question and determined that, in practice, the location of the boundaries do not matter much; there are about as many in the high performance range whether "top" is defined as the top 20% or the top 33%. We defined as "top" the region 75% to 100% performance, "bottom" as the region 0% to 25%, and "middle" as the wider region 25% to 75%. Figure 5 shows these regions using the example data of Figure 4.

When performance is consistent across tests and validating activities, the three performance fractions will be more or less horizontal, they will fall within the same performance categories. We hoped to validate each subskill by demonstrating that individual students did perform consistently across tests and games. It is clear from both Figures 4 and 5 that the performance of child #9007 on pretest, game, and posttest meets neither the precondition of dichotomous performance nor the criterion for consistent performance that would tend to validate the test for this subskill.

VALIDATION CATEGORIES

Stated in terms of our categories of performance derived above, a validation case consists of a triple of performances on pretest, validation activity, and posttest all three of which lie within a single region: "top" or "middle" or "bottom." In this section, we examine possible results in which at least one of the triplet of performances lies in a region different from the other two. These results are not validating.

What triplet of performance will tend to invalidate the test for a particular child and subtask? Generally there are two distinguishable classes of performance profiles that we classify as invalidating. In the first one the performance is in a higher region on both the pretest and the posttest than it is on the validating activity. There are five such profiles, shown in Figure 6. Assuming that the validating activity correctly represents "real measurement," the test would yield a false positive for these children on these subtasks. Because thse profiles have the general shape of a Roman vee, we call them "Invalid Vee."

The other class of profiles which we consider invalidating are those in which performance on the pretest and the posttest are both in a lower performance region than on the validating activity. All five such profiles are shown in Figure 7. Because these have the general shape of a capital Greek lambda, we call these cases "Invalid Lambda."

In categorizing validation results, we have dealt so far with three validating profiles (validating top, validating middle, and validating bottom) and ten invalidating profiles (five invalidating vees and five invalidating lambdas). In a world specifically constructed to make life easy for test-makeres, these would be the only categories that exist. Unfortunately, in terms of our performance triples, there are fourteen other possible cases. These fourteen cases divide naturally into two classes of profiles. There are some cildren whose performance generally improves during the validation procedure. All such profiles are shown in Figure 8. Alternatively, the performance of a few cildren generally decline during validation procedure. All such profiles are shown in Figure 9.

What significance do these fourteen profiles have for our decision about the validity of a test? A common feature of all fourteen of these profiles is that performance on the posttest is different from performance on the pretest. This raises two primary possibilities: either the pretest and the posttest are not equivalent or "something happened" during the validation process to change performance. In Section VI we examine the question of test equivalence. Because we were watching carefully and in detail while children performed the pretest, validating activity, and posttest, we were in many cases able to document "what happened" between pretest and posttest. First of all, of course, children learn from the activities themselves or from other children, thus improving their performance from pretest to posttest. Sometimes they are influenced by other students in the validation setting to change in midstream from a correct to an incorrect strategy, so that their performance actually declines from pretest to posttest. Because the validation process went on for an hour, fatigue is also a factor. Because validation took place alongside regular classroom activity, distraction is unavoidable. Finally, there is an irreducible inconsistency of performance that occurs, particularly in the absence of feedback, as children who are not sure about how to do something try several different strategies.

The fourteen profiles just described carry an ambiguous message about the validity of tests, particularly because the performance on the tests themselves is inconsistent from pretest to posttest. Although we can, by other means, demonstrate the equivalence of the tests themselves, for most children there is no way to distinguish between simple instability of performance and the influences on performance of the test validating procedures. We call these cases "neutral"; those shown in Figure 8, which are generally rising, we call "neutral up," while those in Figure 9, which are generally decreasing, we call "neutral down."

THE VALIDATION CUBE

We have examined twenty-seven possible performance profiles that categorize validation results. Each profile consists of a triplet of categories: top, middle, or bottom for each of the performances on the pretest, validation activity, and posttest. Each could, therefore, be represented by a triplet such as (B,M,T) which, for example, would mean bottom performance on the pretest, middle performance on the validating activity, and top performance on the posttest. All 27 triplets can be represented by a 27-celled cube, as shown in Figure 10, where we have presented performance categories on the three validating steps according to the conventional right-handed x, y, z coordinate system. "Bottom" performance is placed nearest to the origin of each axis. Because we are classifying rather than quantifying, the "middle" region is depicted with the same dimensions as the other two.

We call this display of validation results the "validation cube."

Each of the twenty-seven cells in the validation cube corresponds to a single profile described in the previous section. For example, the performance of child #9007 shown in Figure 5 would be classified in the cell labeled "A" in Figure 10 because the child performed at the bottom of the pretest, at the middle on the game, at the bottom on the posttest.

The validation cube can be exploded as in Figure 11 to show the sets of boxes corresponding to the validating, invalidating, and neutral profiles.

Since it is difficult to visualize the validation cube in three dimensions, we slice it for presentation on a page as shown in Figures 12 and 13. In the latter figure, the capital letters V, I, N, stand for the profiles which tend to validate, invalidate, and are neutral respectively. There are only three boxes which are validating, and these are outlined with bold lines.

The initial area test, described in Section IV above, was validated with 52 children who can be described in the following ways: 29 were male and 23 were female; 12 were Black, 16 were Hispanic, 22 were White, and 2 were "other"; 2 were 7 years old, 16 were 8 years old, 19 were 9 years old, 10 were 10 years old, 2 were 11 years old, 2 were 12 years old, and 1 was 13 years old. The validation results are shown in Figure 14 for the two subskills described earlier as "whole units" and "half units." What do these results say about the validity of the test? We feel they constitute a prima facie case that the test is valid for two subtasks of identifying the attribute, using the overall criteria: a large fraction of cases shown dichotomous and consistent performance.

IV TECHNICAL CONSIDERATIONS

We discuss in this section some technical considerations that cannot be avoided if one wishes to close the loopholes on the prima facie case that our procedure can produce tests valid for assessing performance on subtasks:

- Are the different versions of the same test equivalent?
- What is an adequate sample size for validating a test?
- What constitutes the "presence" or "absence" of a subtask?
- How high is "high" performance and how low is "low"?
- Do validation results provide information about the distinguishability and relative difficulty of subtasks?

Before taking up these questions, we need to discuss one significant detail of low performance on the validation procedure. Because we worked in a wide variety of classrooms, regardless of whether or not instruction had occurred in the topics we were testing, we needed to be sensitive to the students' reactions to our requests for performance on skills they may not have known. We were uncomfortable asking students to work for an hour on something they could not do.

The procedure we adopted was as follows: we encouraged all children to try for as long as they could. When a student said he or she could not do a task, we explained the examples on the test as clearly as we could without teaching and then asked them to look at the test items. If at that point they said they could not do it and the staff member felt confident that this was the case, we stopped. For example, there were 7-year-olds who told us they could not tell time except for the "o'clocks and the thirties." Children stopped at various points during the validating procedures.

Our policy was that if a child could not do a subtask, we classified it as a valid bottom performance. This may be criticized as including in the validation results children who did not complete the entire validation process. However, we observed that these children could not, and they said they could not, carry through this procedure. All available evidence pointed to consistent performance at a low level. It was not feasible for all tests to locate a large number of children who could not carry out the tasks and were willing to spend an hour attempting things they could not do.

For each sbtask on the validation charts in Section V, the number of children with whom we had to deal in this way is indicated by the phrase "aborted valid bottoms."

TEST EQUIVALENCE

One product of our test development method is a set of rules for generating each item in alternative versions. Typically we produced alternative versions of each test for the validation process. The validation itself depended crucially on the equivalence of these versions, since its major criterion was consistent performance across similar tasks. The availability of equivalent forms of each test makes pretesting and posttesting possible during validation and secrecy unnecessary in later use. But we do need to demonstrate that the decisions made about a student's performance will not depend on the form of the test administered.

We demonstrated equivalence by giving pairs of tests to a group of students on the same day and comparing the number of errors on each subtask. Parallel forms should yield consistent performance for each student for each subtask. During the development phase, inconsistencies helped us to pinpoint individual items that needed revision. By repeatedly revising our items in response to classroom results, we were able to achieve a high consistency of performance for every subtask on parallel forms of each test.

What is a criterion for "consistent performance"? On each test there were between two and nine opportunities to demonstrate each subtask, with three and four opportunities dominating. For those subtasks with only two opportunities, we judged equivalence according to whether or not there was an equal number of errors on the first test given compared with the second test given for that subtask. For more than two opportunities, exactly equal numbers of errors for each subtask on each test seemed an unreasonably stringent criterion. For these cases we judged equivalence according to whether or not the number of errors on the first test given differed by no more than one from the number of errors on the second test given for that subtask.

Figure 21 shows, by example, how we display equivalency data for the "whole units" subtask of the initial area test. There are four opportunities to demonstrate this subskill on each test. A total of twenty students from the third and fourth grade took the A and B versions of this test. For some, the first test was version A; for others the first test was version B. The number of errors on the first test taken are plotted on the horizontal axis, and the number of errors on the second test taken are plotted on the vertical axis of Figure 21. The number in the cells are the total numbers of students whose pereformance fell in that region. The band outlined boldly shows the boundaries of our criterion for equivalent performance. (An important characteristic of this test of equivalence is the range of performance frm 0 errors to 4 errors on each test.)

Table I shows equivalency results expressed as a percentage for the subtasks on all the tests. Percentage of equivalent performance is equal to (# of cases of equivalent performance)/[(# of cases of equivalent performance)+(# of cases of non-equivalent performance)]x100%.

ADEQUACY OF SAMPLE SIZE FOR VALIDATION

There are several approaches to selecting sample size for studies of people. At one extreme is the case study method, where close attention is paid to individuals, and conclusions are drawn on the basis of small numbers of cases. At the other extreme is the statistical analysis of data from large numbers of people. Our method lies between these extremes, with at least 40 students paticipating in the validation of each test. The maximum validation sample was 79.

We are trying to make a prima facie case for the validity of our tests based on the "validation cube" displays presented in Section V. We feel tha a severe test for the adequacy of the sample size is to cut this number in half, using random selection, and see if the reduced sample still implies validity. Figure 22 shows the results of such a procedure for the two subtasks of the initial area test. The "uncut" data were presented above in Figure 14. For comparison, the half-sample results have been multiplied by 2 and entered in each cell in parentheses in Figure 22. Our feeling is that analysis of the subset would provide as powerful a case for the prima facie validity of this test as did the original full sample size.

We have carried out the above procedure for every subtask of every measurement test. It is cumbersome to show all of these secondary validation cubes. As a rough measure of the confidence in validity, we have defined a "validating percentage" as the fraction (# of validating cases) /[(# of validating cases) + (# of invalidating cases)] converted to a percentage. Table 2 compares the validating percentages for the full sample for each of the 23 subtasks on our tests with the validating percentages for the "half-samples." We feel these results justify the conclusion that the sample size we have chosen is sufficiently large to demonstrate validity, again with the exception of the weight test.

DISTINGUISHABILITY AND RELATIVE DIFFICULTY OF SUBTASKS

The subtasks for which the final versions of our tests are validated are selected by applying our measurement model to the particular task being examined and are refined so that most children perform either "very high" or "very low," with few in the middle for that subtask. Our measurement model is task-oriented and does not incorporate a theory that accounts for differences in performance on subtasks. However, the validation results can be used to provide evidence about the distinguishability between subtasks and their relative difficulty. If all children performed equaly well on all subtasks, one might worry about whether these subtasks had been adequately distinguished from one another and whether independent subtasks do, in fact, exist.

From our validation data we defined a performance percentage for each subtask as the fraction (# of valid top cases)/(total # of valid cases) converted to a percentage. The results for each subtask are shown on Table 3. The differences on performance percentages between different subtasks on each test provide evidence for distinguishability between subtasks, and seem to confirm common-sense notions about relative difficulty of these subtasks.

CRITERIA FOR HIGH AND LOW PERFORMANCE

Our validation depended upon categorizing performance as "top," "middle," or "bottom." It was important to our results that most performances fell in either the "top" or "bottom" categories. How much are our results affected by the location of the boundaries which we place on "high" and "low" performance?

We need to test the sensitivity of the number of validating cases to the location of the boundaries on our performance categories. We tested this sensitivity by analyzing the same data with three sets of boundaries. These boundaries are shown in Figure 2. The results for 71 children who took the extended length measurement test are shown in Table 4.

The first two internal divisions, in which "top" and "bottom" categories are either one-fifth or one-fourth of the region, satisfy our criterion that the "middle" region be the widest. As shown in Table 4, the number of validating cases appears to be insensitive to these two locations of the internal boundaries of performance categories. Even in the radical test of sensitivity that violates our stipulation about the size of the "middle" region, the number of validating cases was changed significantly for only one subtask.

The low sensitivity of validation results to the position of the internal boundaries means that the location of these boundaries may be chosen somewhat arbitrarily. We set the boundaries one quarter of the way from the top and the bottom. This choice yields a middle region twice as wide as the regions at the top and bottom.

Our decision about the location of the internal boundaries that determine "top" and "bottom" performance influenced the number of opportunities we had to include for each subtask on pretest, validation activities, and posttest. We set the minimum number of opportunities at four. This makes it possible for a single error to still be called "top" performance, because it falls on the upper internal boundary.

In analyzing validation data, performance was judged consistent among pretest, validation activities, and posttest if all three points lay within a single region. The internal boundaries were considered parts of both adjacent regions. For example, a performance percentage triplet 75%, 100%, 75%, was considered to be "valid top" whereas a triplet 75%, 50%, 75% was considered to be a "valid middle."

CONCLUSION

In summary, our categorical validation method can be outlined in four steps through which one cycles until success or failure is manifest:

1. Develop a model for the task being probed;

2. Use the model to analyze the task into subtasks;

3. Use games and other "validation activities" to determine that performance on subtasks is "all-or-none;"

4. Develop tests and validate them by showing consistency of "all-or-none" performance between test and validating activities for each subtask.

When successful, this procedure results in tests that characterize and help to diagnose performance without applying a numerical scale to individuals.

When unsuccessful, the procedure can reveal inadequate understanding by the test developer of the task being probed or the appropriate decomposition into subtasks. Repeatedly the procedure has helped to correct our analysis of physical measurement and the ways in which students carry it out.

Lack of success in the procedure can also imply a limitation in the procedure itself. Human performance is complex and we are accustomed to having nature, especially human nature, escape the models we devise to describe it. We hope that this procedure can be adapted to apply to a range of tasks that are important in schools and useful for children.

XI BIBLIOGRAPHY

Alkin, M. C., "Criterion-Referenced Measurement and Other Such Terms." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT, C. W. Harris, M. C. Alkin, and W. J. Popham (eds.). CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Baker, Robert L., "Measurement Considerations in Instructional Project Development." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT, Chester W. Harris, Marvin C. Alkin, and W. James Popham (eds.) CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Brown, John Seely and Burton, Richard R., "Diagnostic Models for Procedural Bugs in Basic Mathematical Skills." COGNITIVE SCIENCE, 2, 155-192 (1978).

Campbell, Donald T. and Fiske, Donald W., "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." PSYCHOLOGICAL BULLETIN, 1959, 56, 81-105.

Hambleton, R. K., Swaminathan H., Algina J. and Douglas B. C. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments." Review of Educational Research, Winter 1978, Vol. 48, No. 1, pp. 1-47.

Harris, Chester W., "Problems of Objectives-Based Measurement." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT, Chester W. Harris, Marvin C. Alkin, and W. James Popham, (eds.). CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hoffman, Banesh, THE TYRANNY OF TESTING. New York: Collier Books, 1967. (soft cover; 223 pages).

Houts, Paul L., THE MYTH OF MEASURABILITY. New York: Hart Publishing Company, Inc., 1977.

Keesling, J. Ward, "Empirical Validation of Criterion-Referenced Measures." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT, Chester W. Harris, Marvin C. Alkin, and W. James Popham, (eds.). CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Nitko, Anthony J., "Problems in the Development of Criterion-Referenced Tests: The IPI Pittsburgh Experience." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT, Chester W. Harris, Marvin C. Alkin, and W. James Popham (eds.). CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Popham, W. James, "Selecting Objectives and Generating Test Items for Objectives-Based Tests." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT, Chester W. Harris, Marvin C. Alkin, and W. James Popham (eds.). CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Popham, W. James, CRITERION-REFERENCED MEASUREMENT, Englewood, New Jersey: Prentice-Hall, Inc., 1978.

Schwartz, Judah L. and Taylor, Edwin F., "Valid Assessment of Complex Behavior: The TORQUE Approach." THE QUARTERLY NEWSLETTER OF THE INSTITUTE FOR COMPARATIVE HUMAN DEVELOPMENT, July 1978, Volume 2, No. 3.

Skager, Rodney W., "Generating Criterion-Referenced Tests from Objectives-Based Assessment Systems: Unslved Problems in Test Development, Assembly, and Interpretation." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Wilson, H. A., "A Judgmental Approach to Criterion-Referenced Testing." In PROBLEMS IN CRITERION-REFERENCED MEASUREMENT. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Zacharias, Jerrold R., "The Trouble with IQ Tests." In THE MYTH OF MEASURABILITY, Paul L. Houts (ed.). New York, Hart Publishing Company, 1977.

APPENDIX:

SUBTASKS FOR MEASUREMENT TESTS

TIME-TELLING TEST

This test focuses on scale reading: the task of reading a traditional clock face and reporting the time in any of the conventional written or oral notation systems. Students are not asked to measure time intervals. The test has been validated for the following subtasks:

1. Reporting the minute scale for the 1/2 hour, 1/4 hour, and 3/4 hour positions.

2. Reporting the minute scale for the 5 minute positions.

3. Reporting the minute scale for the 1 minute positions.

4. Reporting the hour scale, even when the hour hand is between two numbers.

5. Reporting the correct relationship between the minutes and the preceding or approaching hour.

INITIAL LENGTH TEST

This test focuses on scale reading by presenting lengths to be measured with a ten-centimeter ruler which is calibrated to .5cm. This ruler has a blank tab one centimeter in length before the zero point and a blank tab two centimeters in length after the ten centimeter point.

This test has been validated for the following subtasks:

1. Choosing a correct starting point: Placing the ruler correctly along the line to be measured.

2. Measuring lines of integer length which are shorter than the ruler, such as 7 cm.

3. Measuring lines of non-integer length which are shorter than the ruler, such as 7-1/2 cm.

4. Measuring lines of integer length which are longer than the ruler (between 11 cm and 19 cm.)

5. Identifying the "longest" or "shortest" side of a trapezoid and measuring its length.

EXTENDED LENGTH TEST

This test focuses on scale reading and judging appropriate precision. Students measure lines with a ten-centimeter tab ruler calibrated to .1 cm.

The subtasks are:

1. Choosing a correct starting point. Placing the ruler correctly along the line to be measured.

2. Measuring lines of integer length which are shorter than the ruler, such as 7 cm.

3. Measuring lines of non-integer length which are shorter than the ruler, such as 7.3 cm.

4. Measuring lines of integer length which are longer than the ruler (between 11 cm and 19 cm).

5. Identifying the "longest" and "shortest" sides of a triangle and the "length" of a pencil and measuring them.

INITIAL AREA TEST

This test, described in detail in the text of the paper, assesses performance on the task of identifying the attribute of area. The test helps teachers know whether or not a student can distinguish area from the shape in which it occurs and from lengths. Students use a transparent acetate "ruler," composed of a strip of five "tile" units, to cover regions and compute and report area.

The test has been validated for the following subtasks:

1. Measuring the area of rectangular or irregular shapes which have interior cells and which can be covered by whole units. "Interior cells" refers to that surface area which, when covered by unit "tiles," does not lie along an edge.

2. Measuring the area of irregular shapes which have no interior cells but which must be covered in part by rectangular half-tiles.

EXTENDED AREA TEST

This test examines performance on the tasks of identifying, measuring, and reporting the area of a variety of shapes. The student uses a ten-centimeter ruler to measure lengths, from which area can be computed. The nonrectangular shapes on this test defy routine multiplications of "length times width": students must apply their understanding of the formula.

The test has been validated for the following subtasks:

1. Computing the area of a rectangle whose dimensions must be

measured.

2. Computing the area of an irregular shape whose dimensions must be measured.

3. Computing the area of a right triangle whose dimensions must be measured, and which is presented as half of a rectangle.

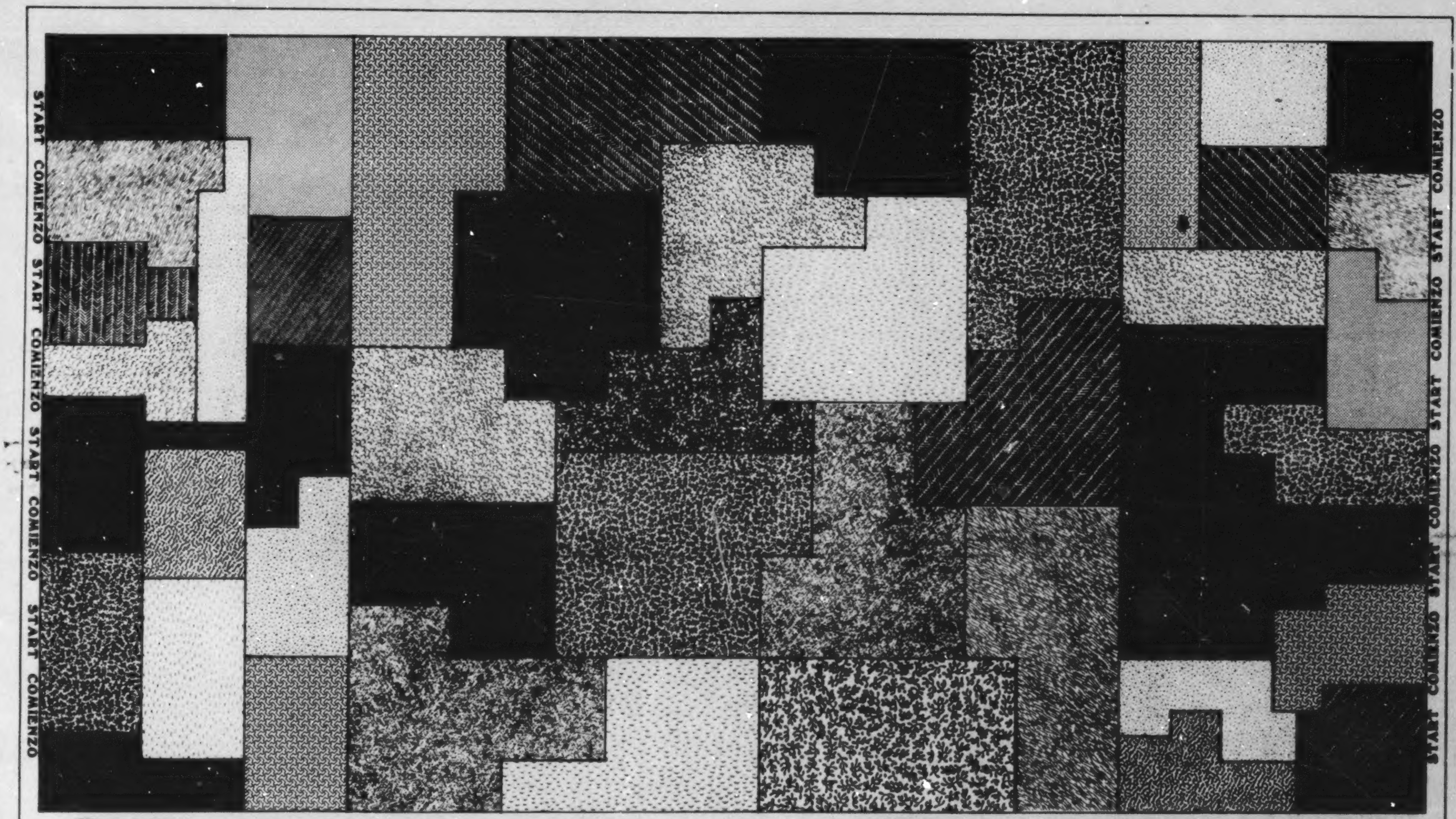
VOLUME TEST

This test asks students to calculate the number of unit cubes in a three-dimensional figure pictured in perspective. Students need to devise strategies other than unit counting in order to find the number of "cubes" needed to construct each building pictured on the test.

This test has been validated for the following subtasks:

1. Finding the number of cubes in a "regular" solid built from unit cubes.

2. Finding the number of unit cubes in an "irregular" solid.


START COMIENZO START COMIENZO START COMIENZO START COMIENZO
START COMIENZO START COMIENZO START COMIENZO START COMIENZO

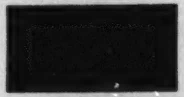
RANCHOS ANCHOS

Name
Nombre ______________________ Grado ____ ate
. echa ______________ ●

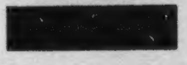
Example
Ejemplo

This shape has an area of 1 tile.
(Check it with your tile-ruler.)
Esta figura tiene un área de 1 teja.
(Mídela con una regla de tejas.)

What is the area of each of these shapes?
¿Cuál es el área de cada una de estas figuras?



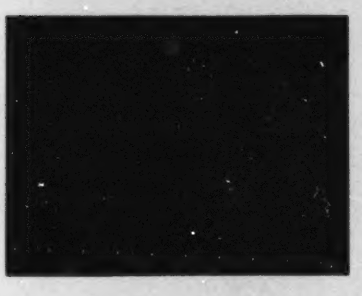
Answer
Respuesta ____________ **tiles**/***tejas***
How many? *¿Cuántos?* What? *¿Qué?*



Answer
Respuesta ____________ **tile**/***teja***
How many? *¿Cuántos?* What? *¿Qué?*

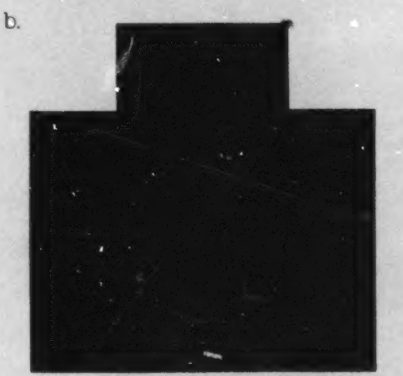
What is the area of each of these shapes?
¿Cuál es el área de cada una de estas figuras?

a.



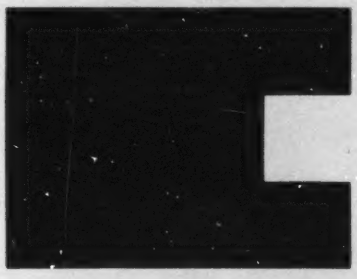
a. Answer
Respuesta ____________ **tiles**/***tejas***
How many? *¿Cuántos?* What? *¿Qué?*

b.



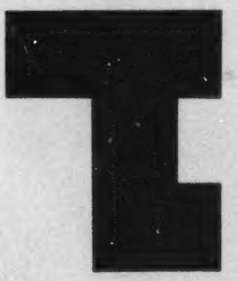
b. Answer
Respuesta ____________ **tiles**/***tejas***
How many? *¿Cuántos?* What? *¿Qué?*

c.



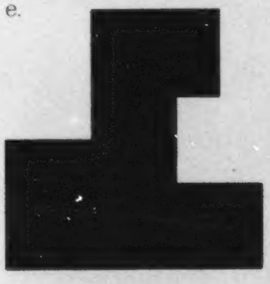
c. Answer
Respuesta ____________ **tiles**/***tejas.***
How many? *¿Cuántos?* What? *¿Qué?*

d.



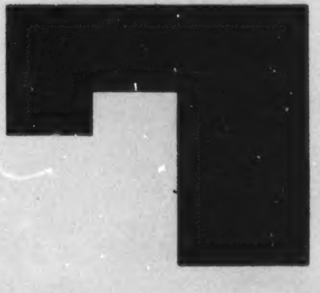
d. Answer
Respuesta ____________ **tiles**/***tejas***
How many? *¿Cuántos?* What? *¿Qué?*

e.



e. Answer
Respuesta ____________ **tiles**/***tejas***
How many? *¿Cuantos?* What? *¿Qué?*

f.



f. Answer
Respuesta ____________ **tiles**/***tejas.***
How many? *¿Cuántos?* What? *¿Qué?*

•edc / TORQUE
copyright © 1979
education development center, inc.
newton mass 02160

Figure 5

Three Performance Categories for Child #9007

Whole Unit Subskill, Initial Area Test

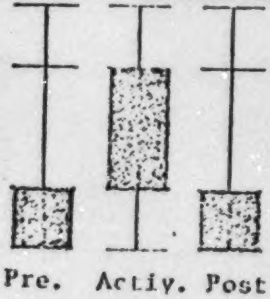


Figure 4

Three Performance Fractions for Child #9007

Whole Unit Subskill, Initial Area Test

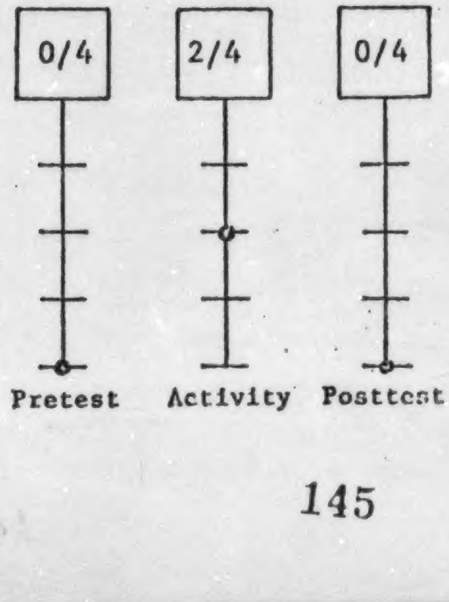
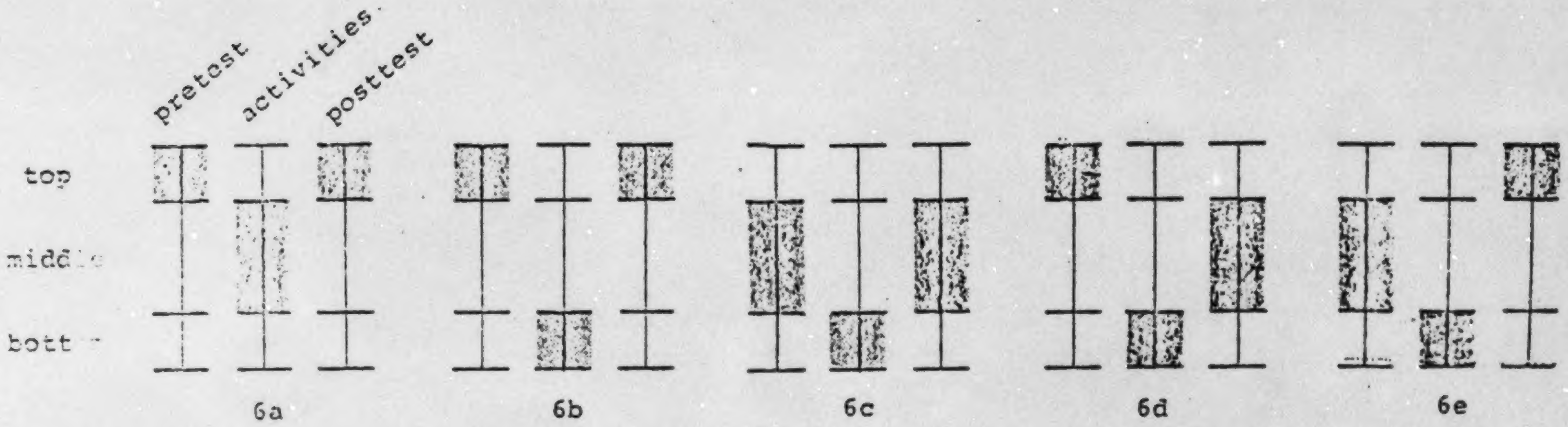


Figure 6

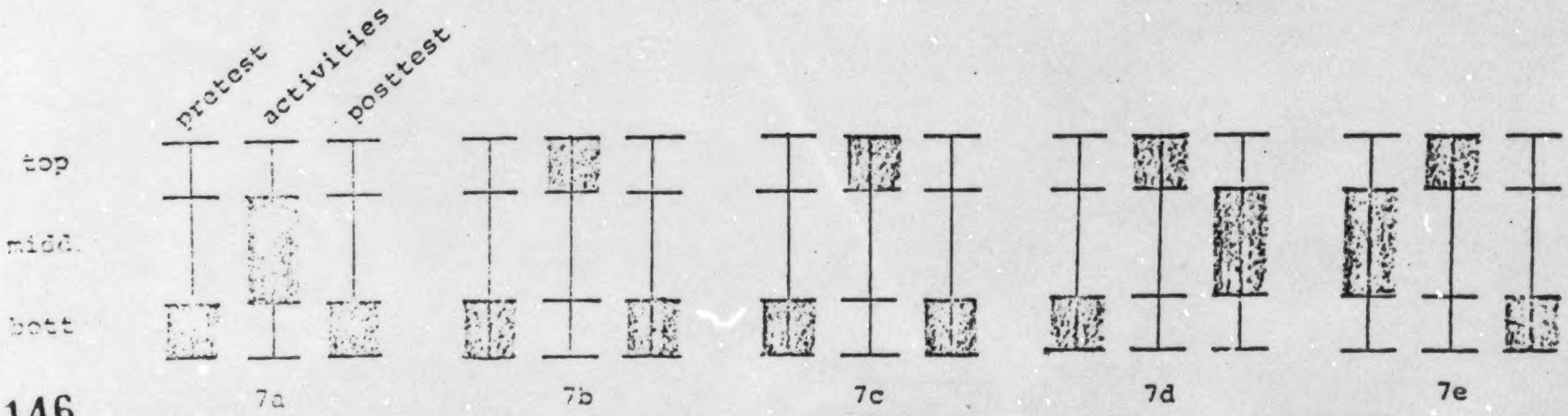
Five Invalidating Profiles That Give Rise to False Positives on Tests - "Invalid Vee"


pretest
activities
posttest
top
middl
bott
6a
6b
6c
6d
6e

Note: For the profiles 6d and 6e, the performance category for the pretest is different from the performance category for the posttest.

Figure 7

Five Invalidating Profiles That Give Rise to False Negatives on Tests - "Invalid Lambda"


pretest
activities
posttest
top
midd
bott
7a
7b
7c
7d
7e

Note: For profiles 7d and 7e, the performance category for the pretest is different from the performance category for the posttest.

Figure 8
Seven Profiles Which Show Improving
Performance - "Neutral up"

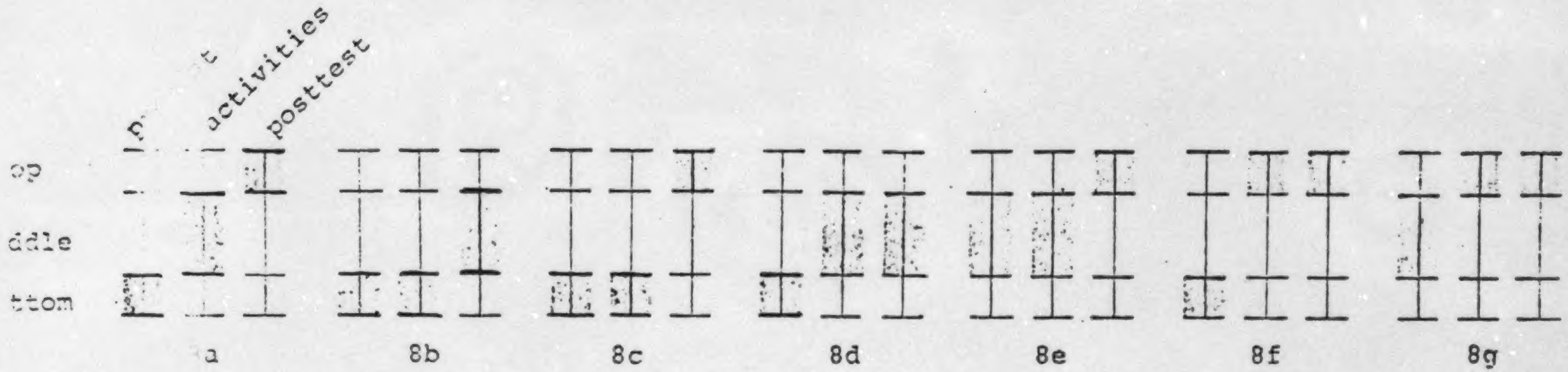


Figure 9
Seven Profiles Which Show Declining
Performance - "Neutral down"

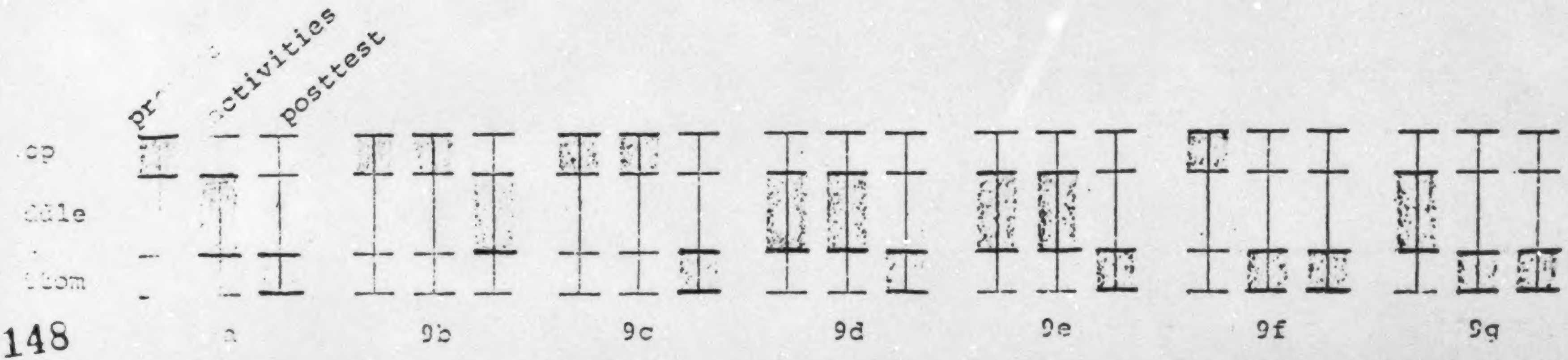
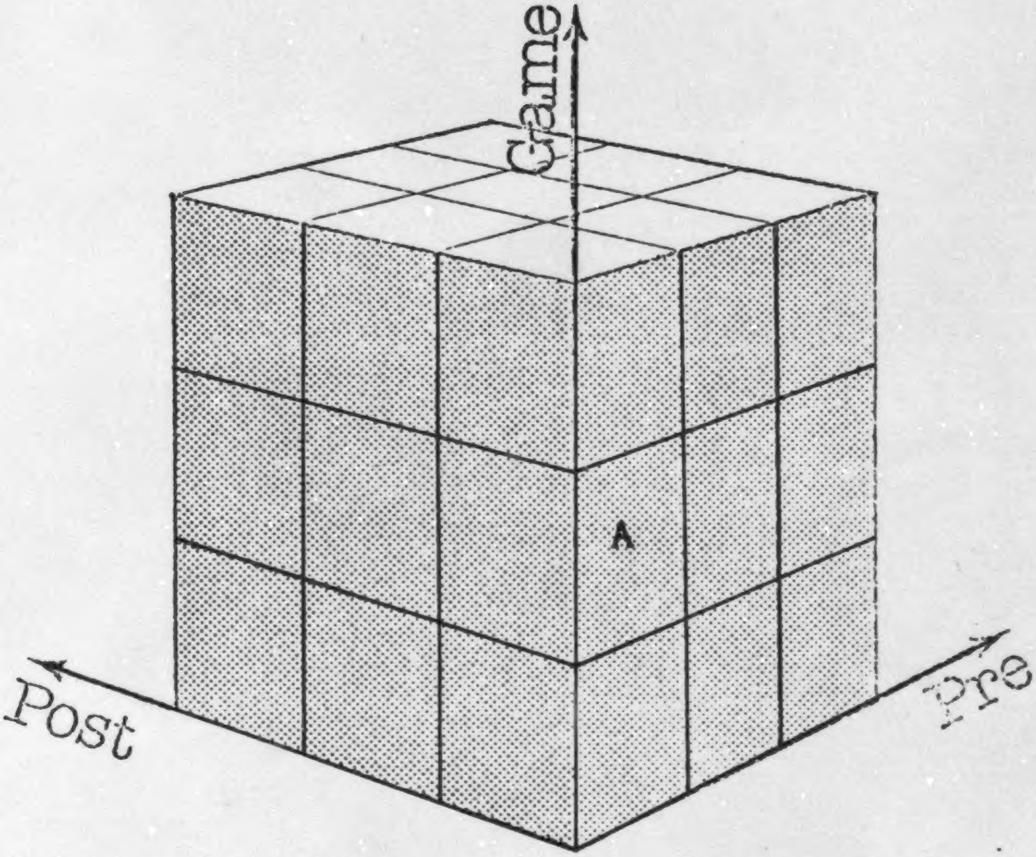


Figure 10 Validation Cube


Game
A
Post
Pre

The box labeled "A" is the location of the performance classified as (B,M,B)

Figure 11 Exploded View of Validation Cube

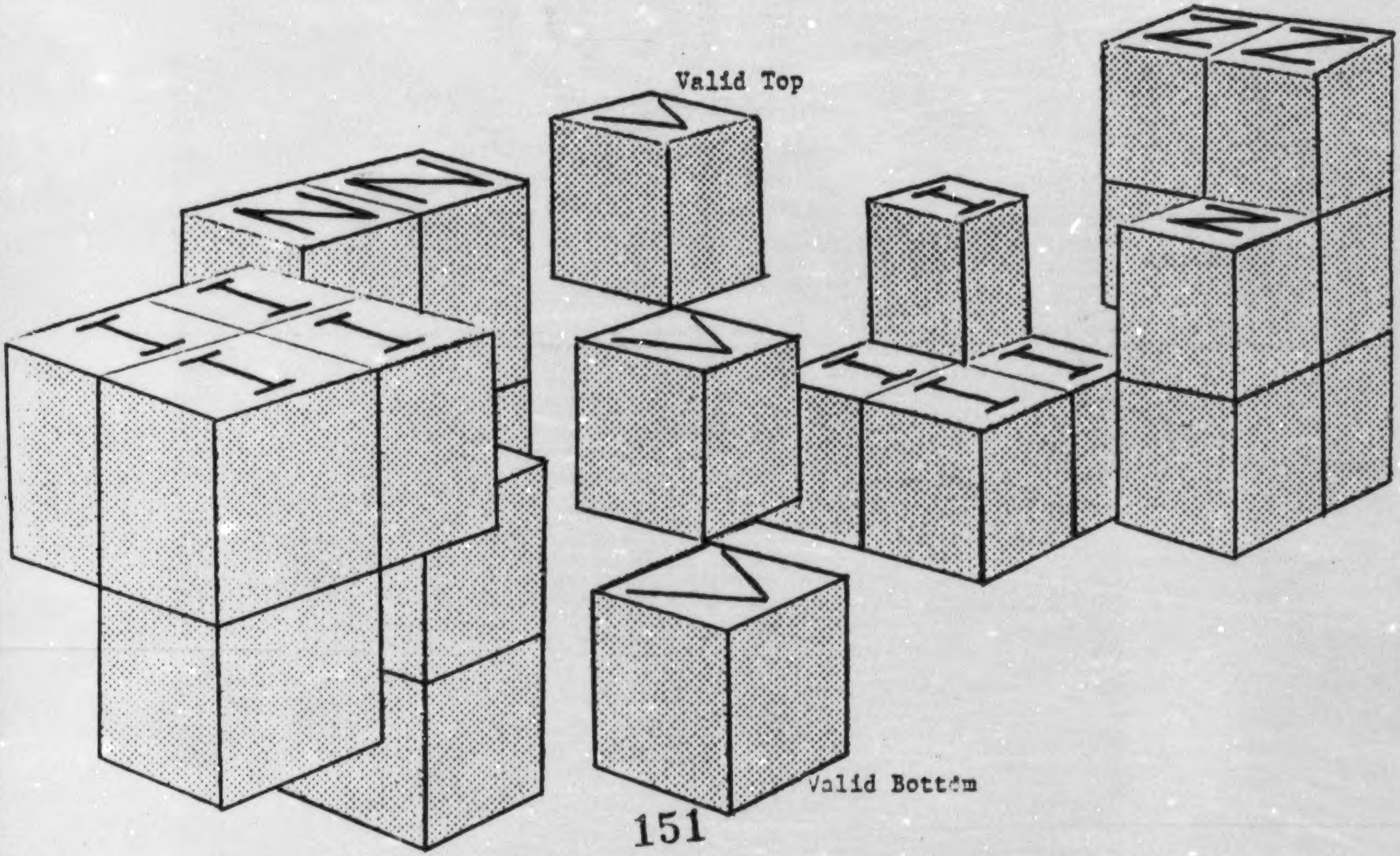

Valid Top
Valid Bottom

Figure 12. Sliced View Of Validation Cube

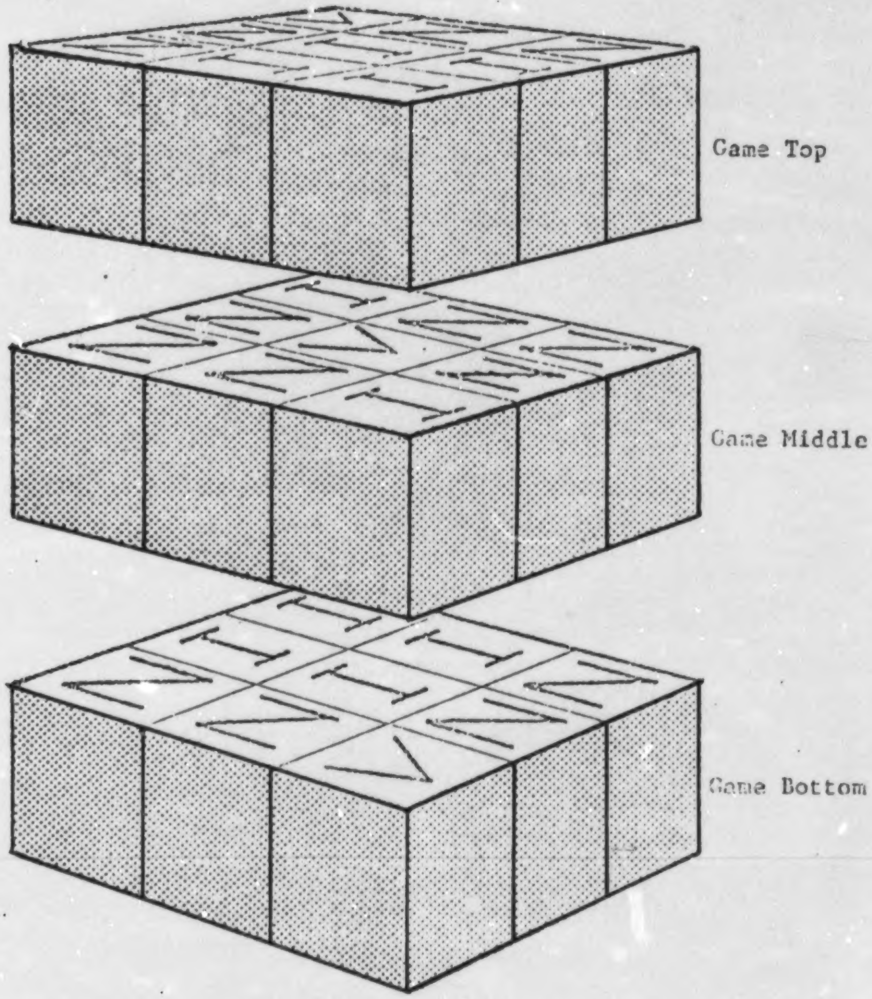

Game Top
Game Middle
Game Bottom

Figure 13. Flat Page Representation of Validation Cube

Game Top

	N	N	**V**
Post	I	I	N
	I	I	N

Pre

Game Middle

	N	N	I
Post	N	**V**	N
	I	N	N

Pre

Game Bottom

	N	I	I
Post	N	I	I
	V	N	N

Pre

Figure 14. Validation Results for Initial Area Test

Whole Unit
"covering"

game top

T	1 N	1 N	14 V	
M	0 I	0 I	2 N	Post
B	0 I	0 I	0 N	
	B	M	T	
		Pre		

game middle

T	0 N	0 N	0 I	
M	0 N	0 V	1 N	Post
B	2 I	0 N	0 N	
	B	M	T	
		Pre		

game bottom

T	0 N	0 I	1 I	
M	2 N	0 I	0 I	Post
B	27 V	1 N	0 N	
	B	M	T	
		Pre		

4 Aborted Valid Bottom

Half Unit
"covering"

game top

T	1 N	2 N	6 V
M	0 I	0 I	1 N
B	0 I	0 I	0 N
	B	M	T
		Pre	

game middle

T	1 N	1 N	0 I
M	0 N	2 V	0 N
B	0 I	0 N	0 N
	B	M	T
		Pre	

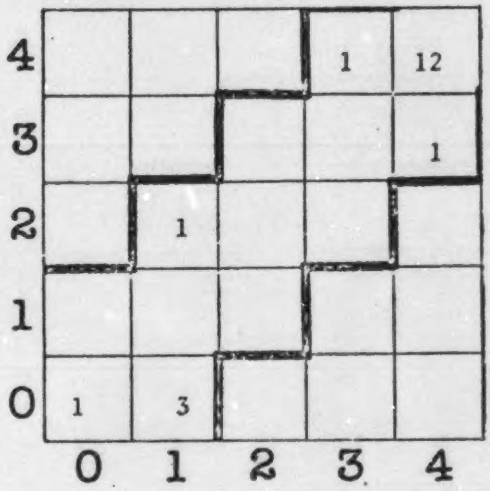
game bottom

T	1 N	0 I	1 I
M	0 N	1 I	0 I
B	35 V	0 N	0 N
	B	M	T
		Pre	

5 Aborted Valid Bottom

Figure 21 Equivalency Results for Half Units Subskill of Initial Area Test

(n = 20)



# of errors on second test taken

# of errors on first test taken

	0	1	2	3	4
4				1	12
3					1
2		1			
1					
0	1	3			

Figure 22. Half Sample Validation Results for Initial Area Test

game top

T	1(0) N	1(2) N	14(10)V
M	I	I	2(2) N
B	I	I	N
	B	M	T
		Pre	

game top

T	1(2) N	2(2) N	6(2) V
M	I	I	1(0) N
B	I	I	N
	B	M	T
		Pre	

game middle

T	N	N	I
M	N	V	1(0)N
B	2(2) I	N	N
	B	M	T
		Pre	

game middle

T	1(0)N	1(0) N	I
M	N	2(2) V	N
B	I	N	N
	B	M	T
		Pre	

game bottom

T	N	I	1(0)I
M	2(2) N	I	I
B	27(34)V	1(0) N	N
	B	M	T
		Pre	

game bottom

T	N	I	1(0)I
M	N	1(0)I	I
B	35(42)V	N	N
	B	M	T
		Pre	

Figure 23. Alternative Boundries for Performance Categories

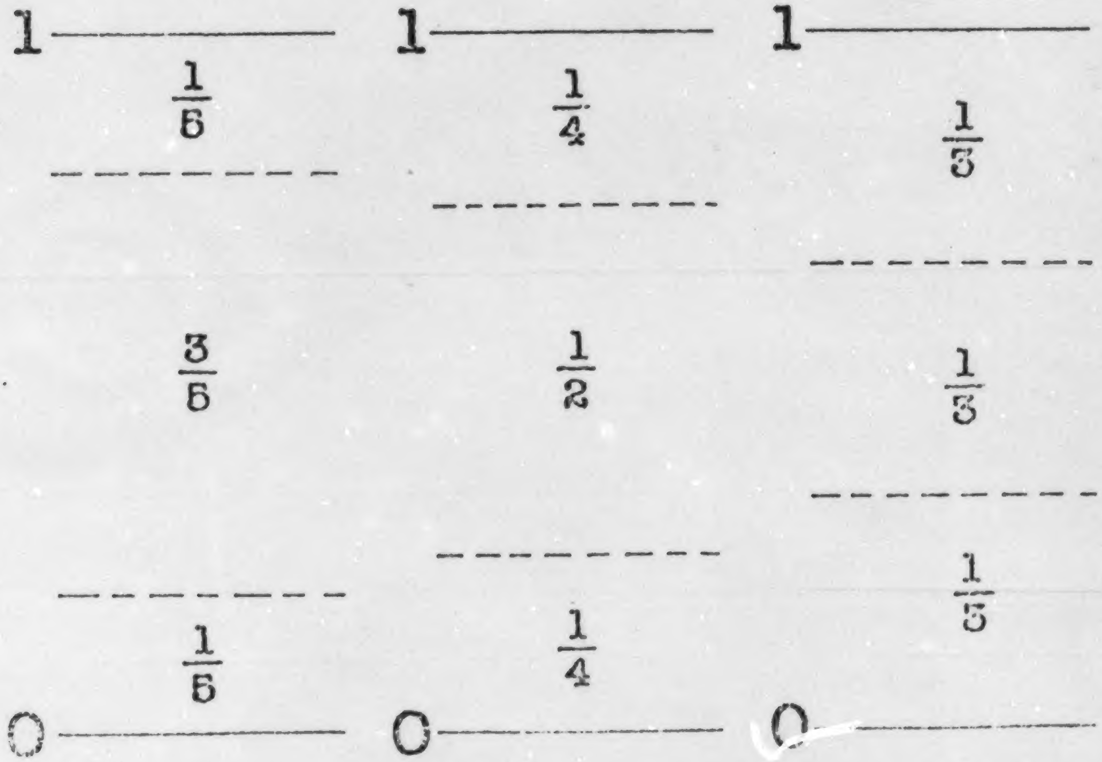

1
1/6
3/6
1/6
0
1
1/4
1/2
1/4
0
1
1/3
1/3
1/3
0

Table 1

Percentage of Dichotomous Performance of Subtasks for Pretest, Validating Activities, and Posttest

Subtasks	Pretest	Validating Activities	Posttest
Time-Telling:			
Minutes to 15	91%	97%	92%
Minutes to 5	93%	89%	89%
Minutes to 1	91%	95%	80%
Hour Scale	62%	76%	78%
Lang./Rel.	96%	96%	97%
Initial Length:			
Starting Point	93%	85%	98%
Attribute	93%	97%	97%
Shorter than the Ruler	100%	100%	98%
Longer than the Ruler	88%	87%	92%
Non-integer	89%	89%	93%
Extended Length:			
Starting Point	97%	94%	95%
Attribute	85%	94%	86%
Shorter than the Ruler	95%	92%	98%
Longer than the Ruler	98%	77%	98%
Non-integer	86%	78%	88%
Initial Area:			
Whole Units	96%	94%	90%
Half Units	90%	92%	94%
Extended Area:			
Rectangular Regions	98%	100%	100%
Irregular Regions	98%	98%	100%
Triangular Regions	96%	96%	98%
Volume:			
Rectangular Solids	86%	89%	95%
Irregular Solids	85%	89%	93%

The appendix gives explanations of the shorthand labels used to describe the subtasks for each measurement test.

Table 2

Percentages of Equivalent Performance

For Each Subskill on Each Test

Time-Telling:	
Minutes to 15	100%
Minures to 5	100%
Minutes to 1	98%
Hour Scale	94%
Lang./Rel.	100%

Initial Length:	
Starting Point	90%
Attribute	86%
Shorter than the ruler	100%
Longer than the ruler	73%
Non-integer	83%

Extended Length:	
Starting Point	86%
Attribute	100%
Shorter than the ruler	100%
Longer than the ruler	95%
Non-integer	62%

Initial Area:	
Whole Units	100%
Half Units	100%

Extended Area:	
Rectangular Regions	92%
Irregular Regions	100%
Triangular Regions	100%

Volume:	
Rectangular Solids	89%
Irregular Solids	95%

Table 3

Percentage of Validating Cases for Each Complete Validation Sample and a Random Half-Sample, for Each Test

Subskills	% Valid for Total Sample	% Valid for Half-Sample
Time-Telling:		
Minutes to 15	92	97
Minutes to 5	97	94
Minutes to 1	90	86
Hour Scale	85	72
Lang./Rel.	99	97
Initial Length:		
Starting Point	88	87
Attribute	100	100
Shorter than the Ruler	98	100
Longer than the Ruler	94	94
Non-integer	88	87
Extended Length:		
Starting Point	100	100
Attribute	83	78
Shorter than the Ruler	95	90
Longer than the Ruler	83	90
Non-integer	88	93
Initial Area:		
Whole Units	93	96
Half Units	96	96
Extended Area:		
Rectangular Regions	98	96
Irregular Regions	98	95
Triangular Regions	100	100
Volume:		
Rectangular Solids	92	100
Irregular Solids	91	92

Table 4

Performance Percentage for Subskills

$$\left(\% = \frac{\#\text{ Valid Top}}{\text{Total }\#\text{ Valid}}\right)$$

Time-Telling:	
Minutes to 15	91%
Minutes to 5	76%
Minutes to 1	72%
Hour Scale	65%
Lang./Rel.	85%
Initial Length:	
Starting Point	80%
Attribute	98%
Shorter than the Ruler	100%
Longer than the Ruler	66%
Non-integer	63%
Extended Length:	
Starting Point	89%
Attribute	100%
Shorter than the Ruler	98%
Longer than the Ruler	69%
Non-integer	14%
Initial Area:	
Whole Units	34%
Half Units	14%
Extended Area:	
Rectangular Regions	34%
Irregular Regions	17%
Triangular Regions	26%
Volume:	
Rectangular Solids	56%
Irregular Solids	50%

Table 5

Sensitivity of Validation Results to Boundaries
of Performance Categories for Extended Length Measurement Test

Subskills	Top & Bottom Fractions	Validating	Neutral	Invalidating
Identify Attribute	1/5	51	7	13
	1/4	54	7	10
	1/3	64	3	4
Starting Point	1/5	57	10	4
	1/4	60	10	1
	1/3	60	10	1
Integer Length Shorter	1/5	64	4	3
	1/4	65	3	3
	1/3	68	1	2
Integer Length Longer	1/5	48	9	14
	1/4	50	9	12
	1/3	51	9	11
Non-integer Length	1/5	50	12	9
	1/4	51	12	8
	1/3	55	10	6

CONCLUSION AND RECOMMENDATIONS

Conventional educational tests, we have argued, do not serve teaching and learning well. There is little evidence that teachers use testing much to guide instruction in the classroom. Yet, at the same time, there is considerable public pressure to increase the amount of testing in the schools.

This widespread public call for an increase in testing provides an unusual opportunity, we think, to begin to develop new assessment practices, more helpful to teachers in the practice of teaching. Assessment, we have argued, should be viewed as an integral part of the teaching and learning process. If this view is correct, there are a number of guidelines that should be followed, in developing new assessment materials.

ASSESSMENT THAT REFLECTS THE CHARACTER OF TEACHING AND LEARNING

In general, we believe the preparation of test materials should begin and end in the classroom, in close interaction with teachers and students. If assessment materials are to serve instruction, they must be informed by an understanding of the ways in which children learn and demonstrate their knowledge in the subject areas assessed. Too often, the only empirical work underlying conventional standardized tests is a statistical analysis of test-item scores performed at the end of the test development process. And, for many objectives-based tests, no empirical work is done at all.

We believe test development should ordinarily involve three loosely-defined steps: open-ended observation of children and their work, the development of somewhat more focused assessment activities, and finally (in some cases) the development of formal assessment instruments. The preparation of test materials should begin with careful observation of children engaged in the sorts of learning tasks the tests are designed to assess. Only by observing children and their work is it possible to identify the kinds of strengths and weaknesses children typicaly display in coming to terms with a subject area.

If observation is successful, it should lead to the development of more focussed games, exercises, and activities that embody the learning tasks being assessed. These games and activities — midway between open-ended observation and formal tests — should elicit some of the patterns and regularities underlying children's work. By watching children completing these games and activities, observers should be able to identify some of the competencies individual children display, differentiate among typical errors, and interpret these errors in terms of the trains of thought that might have produced them.

In some subject areas, we believe, semi-focused games and exercises may be the most rigorous form of assessment desirable. Sometimes — particularly in the sciences, social studies, and the arts — there is no good reason to move from informal exercises to the development of formal tests (other than teacher-made tests). In other areas — particularly reading, writing, and elementary mathematics — formal, easily administered tests may have important instructional value.

When formal tests are developed, we believe they should be based on extensive work with children, based on less formal exercises and activities. Only in this way is it possible to assert with any confidence that the formal tests adequately represent the learning tasks being assessed. Furthermore, we believe that when formal tests are developed, they should not simply be scored in terms of questions right and wrong. Instead, test items should be designed to elicit commonly occuring errors, and the test scoring system should call attention to the precise kinds of errors each child has made.

In most cases, we believe, it is not possible to use multiple choice questions to obtain the sorts of error information needed. Open-ended "constructed answer" questions permit students to make a wide variety of errors, and this diversity is essential in attempting to determine the source of student strengths and weaknesses.

We have focused so far on the development of new assessment materials. While we believe new materials are important, we believe it is equally important to find ways of helping teachers improved their day-to-day skills in observing students and interpreting their work.

One of the central ways in which a teacher can guide a student`s learning is by gaining insight into how a child is thinking in a particular situation, and where the child might usefully move next. The sensitivity and skill involved in this sort of continuing assessment and diagnosis is difficult to acquire, and there is little research to indicate what sorts of training programs might be successful. But we believe additional work in this area could be extremely important.

In addition, new techniques need to be developed to assist teachers in evaluating and interpreting students` regular classroom work. Student essays, art work, problem sets, stories, and projects form a rich source of diagnostic information, much of which regularly goes unharvested. Some techniques that help teachers draw diagnostic information from regular clasroom work have begun to appear, but more work is needed.

Finally, we believe that much can be learned by looking at the development of children`s work over fairly long spans of time — longer than a regular school year. Ways need to be found to collect systematic samples of student work over time, so that teachers can use the work to uncover student strengths, gauge student progress, and discover continuing problems. This approach to assessment -- sometimes called documentation -- has been implemented by Patricia Carini at the Prospect School in Vermont.

ASSESSMENT THAT RESPECTS DIVERSITY

Questions about whether a child has mastered a particular cognitive skill are rarely if ever answered once and for all. A child who can compute the area of a geometric figure in one context, for example, may fail to display the skill at all in another context. A child who speaks fluently in one context may speak only in one or two word sentences in another. And a child who writes in detail on one subject may write haltingly on another.

Children respond differently in different contexts partly as a result of differences in interests and tastes. But partly, as we have argued, these contextual differences are more profound. They arise because children of different cultures bring different stocks of knowledge and experience to bear on cognitive tasks.

If assessment is to serve instruction, it must capitalize on this diversity among cultures and among children within cultures. Assessment materials should offer students multiple contexts in which to demonstrate competence. Thus, for example, diagnostic materials in reading should as a matter of course include a variety of topics and styles, and diagnostic materials in mathematics should include problem sets in widely differing contexts.

Furthermore, we believe test materials should include guidelines for teachers, indicating how cognitive tasks similar in structure to those on the tests can be created using local contexts and materials. Often, we believe, assessment can be strengthened by drawing on stories and topics from the local community, or even the classroom — including materials created by children themselves.

One of the principal elements of the practice of teaching is choosing materials for each child that are likely to engage his skills and competencies. Skills developed in one context can then be strengthened and expanded, so that they can be applied in increasingly diverse and challenging settings. While the development of diverse assessment materials can help in the process of identifying strengths and capitalizing upon them, the ultimate success of this process depends on the sensitivity and insight of the teacher. Here, as before, we think that the development of new materials should be coupled with increased resources for in-service training. Materials by themselves, while always necessary, are never sufficient.

ASSESSMENT THAT ENCOURAGES DIALOGUE

Much attention in standardized educational testing has gone into efforts to express test results as numerical scores. But often, we believe, quantitative test scores hide as much as they reveal. Particularly for purposes of teaching and learning, we believe more can be gained by looking at student questions and answers themselves than by looking at numerical summaries.

We have argued that tests serve instruction by helping teachers interpret the thought processes underlying student work. A teacher's interpretation of a student's work is always tentative and exploratory, and teachers can often gain insight by discussing the work with the student, other teachers, and parents. Assessment materials can often provide particularly well-focused examples of student work, which can serve as a foundation for this sort of dialogue and discussion.

If assessment materials are to serve as a foundation for dialogue between teachers, students, and parents, then student test forms must be returned to students as soon as possible after the tests are completed. Generally, we believe this means that tests for instructional purposes must be marked by the teacher who administers them (or by the students themselves). It is extremely unlikely that tests which must be sent off for centralized scoring can be returned in time to serve instruction.

In addition, we believe, parents can and should contribute to the assessment of their children's work. One way to do this is to have teachers discuss assessment questions and answers with parents. Also, parents should be encouraged to offer assessments of their own, derived from observations of their children at home. We recognize, of course, that parent involvement in education is a goal frequently stated but difficult to achieve. Becoming closely involved in the education of their children is often especially hard for working parents. We believe that well-designed assessment materials can, by providing clear, focussed examples of children's work, can improve the dialogue between teachers, parents, and students.

Finally, we believe that assessment materials can be used to stimulate dialogue among students. Inevitably, as we have argued, different students approach assessment tasks in different ways, and this diversity provides a rich resource for exploration. By encouraging students to share their approaches to a cognitive task, teachers can help students become aware of alternative problem-solving strategies, their advantages and disadvantages. Dialogue, then, may help students increase their repetoire of cognitive skills.

Dialogue among students may also help promote one additional educational goal. By discussing some of the reasons why their answers to assessment questions differ, students may become more reflective about their own thought processes. Students may learn, when confronting a task, how to generate multiple potential solutions and how to assess their quality. In this way, dialogue among students may help them learn to examine their work, raising questions and identifying strengths and weaknesses.

NEXT STEPS

Most of the ideas we have proposed are not, in themselves, new. They have a long and honorable history in the psychology and philosophy of education. But they have not yet played a very strong role in the development of educational assessment.

The development of alternative assessment practices, of the sort we have described will not be easy or inexpensive, but we believe the investment could reap substantial rewards. We propose the following strategies.

First, it seems to us that, in developing new assessment materials, it is worth starting small. It is, in our view, inappropriate to attempt to construct, all at once, tests that completely cover a subject area, such as elementary school mathematics or junior high writing. It is much more valuable to carve out relatively small, well-focused domains in which careful analysis of the cognitive tasks involved and close empirical work with children can be carried out.

Even if this recommendation is followed, however, development costs are likely to be high, a fact made amply clear by the experience of project TORQUE. Moreover, the foundations that supported the development of an alternative to current assessment practice did not support the implementation of that alternative. The situation at the time of this writing is that these new methods and materials sit on a shelf waiting changes of heart, perspective and practice on the part of publishers. We expect that work in reading and writing will be more expensive and more difficult and find even less enthusiastic support among foundations and publishers.

Second, we recommend that the development of new assessment materals ought to be carried out by groups with a strong interest in the content areas being assessed. These groups should be deeply involved in all aspects of the development process — including observation of children, preparation of materials, and validation. They might also be engaged in pilot efforts to implement the materials developed. It is not sufficient to engage subject-matter specialists simply to review test items once they are written. Educators with strong subject-matter backgrounds must be involved throughout.

Third, we recommend that schools interested in adopting new forms of assessment should begin by focusing on a small number of classrooms and subject areas. The temptation is large to attempt to overhaul a school's assessment program in one, swift step, but we believe such an approach is ill-advised. Implementing the sorts of ideas we have proposed should be an iterative process, in which new practices and organizational relationships are slowly developed.

Finally, making the forms of assessment we have suggested work in practice will depend on the sensitivity and ingenuity of teachers. It is unreasonable to ask teachers to be wise and insightful observers of children and their work if the resources to support classroom teaching are meager and classrooms overcrowded. The strategies we have proposd can not be implemented, at least in the short run, without extra resources for in-service training and materials. In the long run, however, the ideas we have proposed might not cost substantially more than present forms of testing, since many of the materials we have suggested would serve both assessment and instruction.

There is a large and growing demand for improved educational assessment in the classroom. We are firm in our belief that appropriate assessment practices are possible. Although the development of new, more useful assessment materials will require an investment of resources, we believe this investment is likely to have a profound and beneficial effect on teaching. Indeed, as we have argued, successful teaching is in large measure a continuing process of inquiry and assessment.